

Online Resource 1: Supplementary Material

Resample or Reroute? Recoverable Stopping Debt Without Identified Action Selection

Teng-Ruei Chen

Krixvon AI, Taipei, Taiwan

Correspondence: ymchen.bi04g@g2.nctu.edu.tw

ORCID: 0000-0001-8995-334X

Machine Learning submission manuscript

Abstract

This supplement records the exact verifier-gated estimands, finite-bank derivations, descriptive matched-information audit, exact-fold exchangeable-actions calibration, deterministic clustered inference, negative self-attacks, draw-position diagnostics, cross-paper data ledger, and SHA-256 artifact bindings supporting the accompanying article. Historical empirical quantities condition on frozen response banks and declared evaluators. A pointwise realized maximum is treated as a clairvoyant descriptive statistic, not as evidence of heterogeneous expected action value. The supplement also gives a stopped-history value decomposition and states the common-population condition required before multiple gates can compose. A separate LiveCodeBench FIT/DEV/sealed-TEST ladder is reported as a support/transport diagnosis rather than a generic negative result about outcome-blind features. A later preregistered BigCodeBench replication enforces a two-sided support gate before any new feature fit or DEV/TEST opening and records an evaluator-randomness amendment that prevents retry-conditioned labels. A separately identified synthetic support-gate stress test reports an author-frozen generative model, the full 20-cell grid, Monte Carlo intervals, and source bindings; it uses no LLM data.

Version relationship. The experimental record is shared with arXiv:2607.08665v3. The historical empirical record is unchanged. This journal supplement additionally reports a separately identified, author-frozen synthetic support-gate stress test.

Scientific revision chronology

The final ordering was learned rather than assumed. Fresh MBPP+ first established the narrow recovery mechanism. An all-episode matched-information audit then appeared to show a realized-maximum advantage without controller value. Subsequent audit established that the maximum is a signal-free function of realized outcomes and that the audited population includes verifier rejections, so it is not a Gate-3 test of the MBPP+ stopping event. The exchangeable reference was added post-audit to calibrate marginal action label symmetry, not to test conditional heterogeneity. Separately, LiveCodeBench had all-zero fitting action advantages, which prevented its controllers from diagnosing feature insufficiency. BigCodeBench converted that lesson into a preregistered support guard and stopped before controller fitting or development/test opening. These corrections narrow the claims; the experiments do not form a completed chain across populations.

1 Evidence roles and frozen scope

The evidence is deliberately stratified. The confirmatory mechanism result uses a hash-defined TEST split of 152 MBPP+ v0.2.0 queries, ten stochastic draws from each of Qwen2.5-14B-Instruct-AWQ and Qwen2.5-7B-Instruct-AWQ, and an EvalPlus 0.3.1 scorer. Base-test passage is policy visible; Base-and-Plus correctness is evaluator only. Llama-3.1-8B-Instruct and NVIDIA-Nemotron-Nano-9B-v2 are post-outcome model-family sensitivities. The historical 11-model archive is an audit and development record, not an independent replication of the fresh MBPP+ result.

The prespecified observability-ladder follow-up is a separate experiment on 342 LiveCodeBench `code_generation_lite` v5+v6 tasks: one pipeline pilot, 136 FIT, 68 DEV, and 137 sealed TEST queries. It holds the Qwen14 start model and the Qwen14-resample/Qwen7-reroute action set fixed while varying only nested stopped-history information. The FIT/DEV controller was frozen before one TEST outcome opening. Its result does not replace the MBPP+ mechanism. The opened-population matched-information audit takes a second action on every episode and is reported separately because it does not condition on the Gate-1 stopping event.

The BigCodeBench v0.1.4 replication is a second, independently generated support experiment. A pinned authority qualifies 1,100 of 1,140 tasks and assigns one pilot, 483 FIT, 211 DEV, and 405 TEST tasks by near-duplicate component. Its gate fails on FIT and therefore leaves all DEV/TEST outcomes unopened; it cannot replace or retrospectively reinterpret the LiveCodeBench TEST result.

The one-query pilot MBPP/100 is excluded from all scientific splits. The reported generation-token count includes the terminal token, whereas decoded text omits it. Numeric draw indices remain fixed batch positions; neither the generation protocol nor the analysis treats them as independent and identically distributed samples.

Table S1: Canonical notation and information boundary. These symbols replace earlier overloaded uses of H , Y , U , D , and S .

Symbol	Boundary	Meaning	Replaces or disambiguates
$\mathcal{H}_{it}$	policy-visible	Observable history/filtration through call t	H_i used informally for history
ω_i	policy-internal	Frozen policy randomization for query i	U_i as randomization
Z_{ij}	policy-visible	Verifier result or declared stopped-history feature vector	Observable signal only
S	policy-visible event	Declared stopping event, such as Base-test acceptance	Avoids conflict with a resample action label
Y_{ij}^0	evaluator-only	Hidden correctness of the accepted or initial response	H_{ij} used as hidden correctness
Y_{ij}^a	evaluator-only	Hidden correctness of the completion produced by action a	Action-specific correctness
$U_{ij}(a)$	evaluator-only	Utility after action a , usually $\max\{Y_{ij}^0, Y_{ij}^a\}$	Y_a used as utility in the theory section
Δ_{ij}	evaluator-only	$U_{ij}(\text{rer}) - U_{ij}(\text{res})$	D_{ij} as action difference
D	design constant	Number of archived draws per query/model	Draw count only
F_{id}	evaluator-only subset	$\mathbf{1}\{Z_{id}^{14} = 1, Y_{id}^{14} = 0\}$; not a runtime stop	False-positive subset of S
$a \in \{\text{res}, \text{rer}\}$	declared action	Resample or reroute	res/rer action labels

2 Verifier-gated recovery estimands

For TEST query i and draw $d \in \{0, \dots, D-1\}$, let Z_{id}^{14} denote visible 14B Base-test passage and let Y_{id}^m denote hidden Base-and-Plus correctness for model m . Define an irreversible false-positive stop by

$$F_{id} = \mathbf{1}\{Z_{id}^{14} = 1, Y_{id}^{14} = 0\}. \quad (\text{S1})$$

The prespecified independent-pairing recovery opportunity is

$$\hat{\Delta}_{\text{VG}} = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{D} \sum_d F_{id} \right) \left(\frac{1}{D} \sum_e Y_{ie}^7 \right). \quad (\text{S2})$$

The fixed-offset sensitivity is

$$\hat{\Delta}_{\text{off}} = \frac{1}{ND} \sum_{i=1}^N \sum_{d=0}^{D-1} F_{id} Y_{i,(d+1) \bmod D}^7. \quad (\text{S3})$$

The prespecified secondary event is a 7B Base-test rejection:

$$\hat{\Delta}_{\text{sec}} = \frac{1}{N} \sum_i \frac{1}{D} \sum_d \mathbf{1}\{Z_{id}^7 = 0\} \left[\frac{1}{D} \sum_e Y_{ie}^{14} - \frac{1}{D-1} \sum_{e \neq d} Y_{ie}^7 \right]. \quad (\text{S4})$$

It compares fixed rerouting to 14B with leave-one-out fixed resampling of 7B after an observable rejection. It is event-weighted, prespecified as underpowered, and is neither a stopping-debt estimand nor a selector. All three equations are evaluator-only finite-bank quantities and

none selects a response at deployment time; only Eqs. (S2) and (S3) measure false-positive-stop recovery mass.

Table S2: Fresh verifier-gated recovery results. Intervals resample 152 query clusters with 10,000 deterministic bootstrap replicates.

Sequence or sensitivity	Estimate (points)	Interval	Status
Qwen14 false-positive stop $\rightarrow$ Qwen7, (S2)	+2.592	[+1.618, +3.664] (95%)	primary PASS
Qwen14 $\rightarrow$ Qwen7 fixed offset, (S3)	+2.500	[+1.513, +3.618] (95%)	sensitivity PASS
After 7B rejection: reroute 14B – leave-one-out resample 7B, Eq. (S4)	+2.882	[+0.931, +5.201] (95%)	prespecified underpowered secondary
Qwen14 $\rightarrow$ Llama-3.1-8B	+2.974	[+1.553, +4.921] (97.5%)	post-outcome
Qwen14 $\rightarrow$ Nemotron Nano 9B	+3.303	[+1.711, +5.461] (97.5%)	post-outcome

The primary point is the exact reduced fraction 197/7600. Nonzero primary recovery occurs in 28 of 152 query clusters; the ten largest clusters account for 60.2% of the total. The concentration is visible in Figure S1 and is why uncertainty is clustered by query rather than by completion.

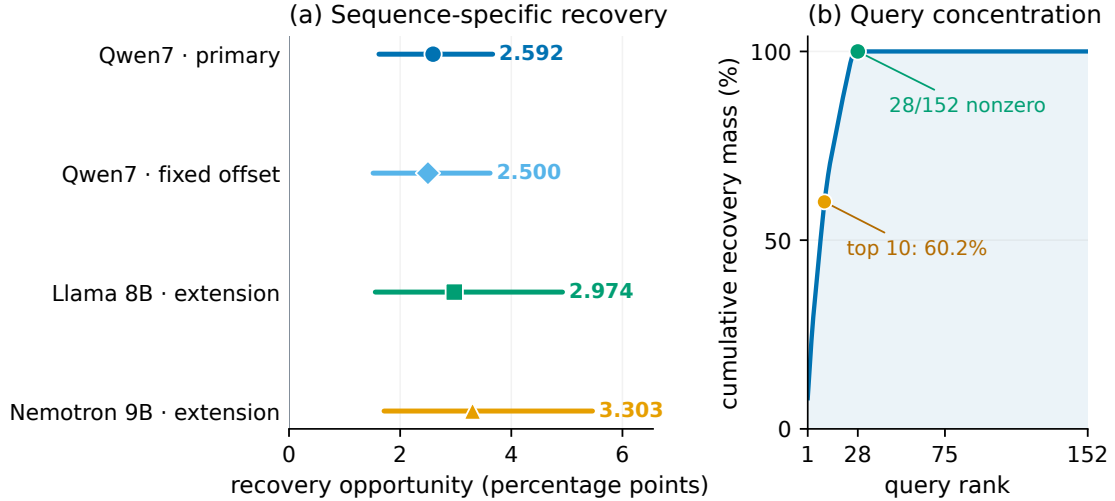


Figure S1: Fresh MBPP+ recovery estimates and query-level concentration. The familywise-corrected extension arms support model-family robustness, not an independent TEST replication.

3 Closed-form finite-bank oracle

For query i and model m , let

$$r_{im} = \min\{d \in \{1, \dots, D\} : Y_{imd} = 1\}, \quad (\text{S5})$$

with $r_{im} = +\infty$ if the fixed-order bank contains no success. If a call to model m costs $c_m > 0$, define

$$C_i^* = \min_m c_m r_{im}. \quad (\text{S6})$$

Proposition 1 (Minimum-cost finite-bank opportunity). *The maximum success indicator attainable by any evaluator-informed legal action sequence under cap B is $\mathbf{1}\{C_i^* \leq B\}$.*

Proof. Reaching the first successful draw of model m requires exactly r_{im} paid calls to that model and therefore cost $c_m r_{im}$. Interleaved calls to other models have nonnegative cost and cannot advance model m 's fixed draw head. Conversely, calling a minimizing model r_{im} times realizes its first success. Minimizing over models proves (S6). $\square$

Proposition 2 (Exchangeable realized-maximum baseline). *If (U_0, U_1) has an exchangeable joint law conditional on S , then*

$$\mathbb{E}[\max\{U_0, U_1\} \mid S] - \max_a \mathbb{E}[U_a \mid S] = \frac{1}{2} \mathbb{E}[|U_1 - U_0| \mid S]. \quad (\text{S7})$$

The gap is positive whenever realized outcomes disagree with positive probability, even though the two conditional expected action values are identical. For independent Bernoulli(p) actions it equals $p(1 - p)$.

Proof. Apply $\max\{x, y\} = (x + y + |x - y|)/2$ and use exchangeability to identify the common conditional mean. In the Bernoulli case, the probability of a disagreement is $2p(1 - p)$. $\square$

Remark 3.1 (General two-action identity). *Exchangeability is not required. Let $\Delta = U_1 - U_0$. For any two integrable actions and any event S with positive probability,*

$$\mathbb{E}[\max\{U_0, U_1\} \mid S] - \max_{a \in \{0,1\}} \mathbb{E}[U_a \mid S] = \frac{1}{2} \mathbb{E}[|\Delta| \mid S] - \frac{1}{2} |\mathbb{E}[\Delta \mid S]|. \quad (\text{S8})$$

The exchangeable value is therefore the largest expected gap compatible with a fixed disagreement rate. A shortfall records a marginal action difference, not observable conditional heterogeneity; the statistic contains no stopped-history signal.

Thus V_{or} is a clairvoyant realization value, not a test of conditional-mean heterogeneity. Empirical action and oracle maxima are also selected statistics; an empirical difference requires an explicit reference distribution before it can support a heterogeneity claim.

The implementation was compared exhaustively with recursive enumeration for all 2^6 binary outcome banks of two models and three draws. The two-Qwen frontier has a maximum oracle–best-fixed difference of 3.289 percentage points at a nominal cap of 280 tenths (the cost of two Qwen2.5-14B calls under the frozen 70/140 nominal costs). The exploratory four-model equal-cost frontier has a maximum difference of 8.553 points. These gaps show opportunity, not an observable selection rule.

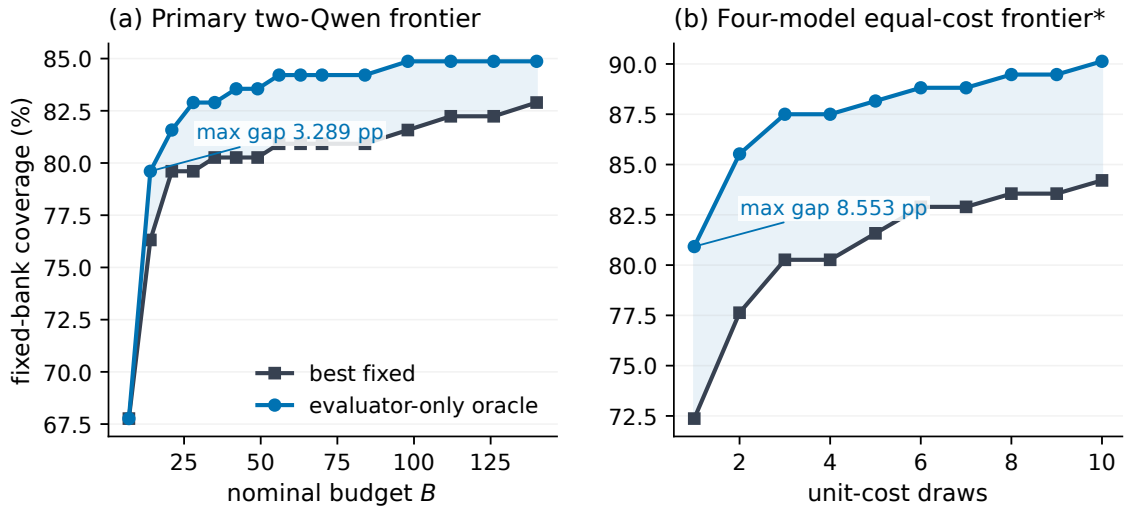


Figure S2: Closed-form finite-bank opportunity frontiers. Shading is the evaluator-only oracle gap above the best fixed model–draw-count policy.

4 Stopped-history observability decomposition

Let S be an observable stopping event with positive probability. Conditional on S , suppose two legal second actions have evaluator utilities $U_0, U_1 \in [0, 1]$, and let Z be any outcome-blind signal measurable from the stopped history. Hidden correctness used to construct U_0 and U_1 is not part of Z . Let $a_{\text{FIT}} \in \{0, 1\}$ be the fixed action frozen from FIT, write $\Delta = U_1 - U_0$, and define

$$V_{\text{frozen}} = \mathbb{E}[U_{a_{\text{FIT}}} \mid S], \quad (\text{S9})$$

$$V_{\text{fix}} = \max_{a \in \{0, 1\}} \mathbb{E}[U_a \mid S], \quad (\text{S10})$$

$$V_Z = \mathbb{E} \left[\max_{a \in \{0, 1\}} \mathbb{E}[U_a \mid Z, S] \mid S \right], \quad (\text{S11})$$

$$V_{\text{or}} = \mathbb{E}[\max\{U_0, U_1\} \mid S]. \quad (\text{S12})$$

These are, respectively, the value of the action selected and frozen on FIT, the best fixed second action in the target population, the Bayes value attainable from the declared stopped-history signal, and the evaluator-only oracle. Distinguishing the first two matters whenever the action ranking does not transport from FIT to the target population.

Proposition 3 (Support-aware stopped-history value decomposition). *The four values satisfy*

$$V_{\text{frozen}} \leq V_{\text{fix}} \leq V_Z \leq V_{\text{or}}, \quad (\text{S13})$$

and the oracle–frozen gap decomposes exactly as

$$V_{\text{or}} - V_{\text{frozen}} = \underbrace{V_{\text{fix}} - V_{\text{frozen}}}_{\text{fixed-action transport loss}} + \underbrace{V_Z - V_{\text{fix}}}_{\text{observable gain}} + \underbrace{V_{\text{or}} - V_Z}_{\text{residual information debt}}. \quad (\text{S14})$$

If Z_1 is a measurable coarsening of Z_2 , then $V_{Z_1} \leq V_{Z_2}$. Thus a nested observability ladder is monotone at the Bayes-value level.

Proof. Because a_{FIT} is one member of the legal action set, $V_{\text{frozen}} \leq V_{\text{fix}}$. For the remaining inequalities, using $\max\{x, y\} = x + (y - x)_+$ gives the identities

$$V_{\text{fix}} = \mathbb{E}[U_0 \mid S] + (\mathbb{E}[\Delta \mid S])_+, \quad (\text{S15})$$

$$V_Z = \mathbb{E}[U_0 \mid S] + \mathbb{E}[(\mathbb{E}[\Delta \mid Z, S])_+ \mid S], \quad (\text{S16})$$

$$V_{\text{or}} = \mathbb{E}[U_0 \mid S] + \mathbb{E}[\Delta_+ \mid S]. \quad (\text{S17})$$

Because $x \mapsto x_+$ is convex, conditional Jensen and the tower property give

$$(\mathbb{E}[\Delta \mid S])_+ \leq \mathbb{E}[(\mathbb{E}[\Delta \mid Z, S])_+ \mid S] \leq \mathbb{E}[\Delta_+ \mid S]. \quad (\text{S18})$$

Subtracting the first identity proves (S13); adding and subtracting V_{fix} and V_Z proves (S14). If Z_1 is a coarsening of Z_2 , apply conditional Jensen once more to $\mathbb{E}[\Delta \mid Z_1, S] = \mathbb{E}[\mathbb{E}[\Delta \mid Z_2, S] \mid Z_1, S]$. $\square$

If second-action utility preserves an already correct stop, $U_a = \max\{Y^0, Y^a\}$, let $V_{\text{stop}} = \mathbb{E}[Y^0 \mid S]$. Then $V_{\text{stop}} \leq V_{\text{frozen}}$ and

$$V_{\text{or}} - V_{\text{stop}} = (V_{\text{frozen}} - V_{\text{stop}}) + (V_{\text{fix}} - V_{\text{frozen}}) + (V_Z - V_{\text{fix}}) + (V_{\text{or}} - V_Z). \quad (\text{S19})$$

The first term is recovery under the FIT-frozen second action; the remaining terms separate action-ranking transport, observable value, and residual evaluator-only information.

When $V_{\text{or}} > V_{\text{fix}}$, the normalized recoverable share

$$\rho_Z = \frac{V_Z - V_{\text{fix}}}{V_{\text{or}} - V_{\text{fix}}} \in [0, 1] \quad (\text{S20})$$

states how much oracle advantage above the target-population best fixed action is recoverable from signal Z . It does not include $V_{\text{fix}} - V_{\text{frozen}}$, which reflects fixed-action selection or split transport rather than stopped-history observability. The primary +2.592-point mechanism estimate in the article is a related event-weighted debt quantity for one prespecified action; it is not itself ρ_Z . Conversely, the cross-fitted ridge results estimate the values of two particular learned policies. They are empirical witnesses below the Bayes envelope V_Z , not proof that $V_Z = V_{\text{fix}}$ for every legal outcome-blind learner.

For LiveCodeBench, the observed TEST values are 0.585714 for stop-only, 0.595238 for FIT-selected resampling, 0.611905 for the descriptive TEST-best fixed reroute action, and 0.621429 for the realized maximum. The associated 1.667- and 0.952-point differences are descriptive selected-sample arithmetic, not estimators of the population terms in Eq. (S14). Because FIT has zero action-difference support, the ladder does not identify V_Z or ρ_Z . The MBPP+ matched audit cannot instantiate this decomposition for Gate 1 because its implicit event is $S = \Omega$, rather than the observable stopping event. The exact stopped/non-stopped population accounting appears in Section 8.

The same value ordering extends to any finite action set by replacing the two-action maximum in (S10)–(S12) with $\max_{a \in \mathcal{A}}$. If $\mathcal{A}_1 \subseteq \mathcal{A}_2$, the absolute fixed, signal-informed, and oracle values cannot decrease. Their *differences*, however, need not increase. For example, under two equally likely values of an observed Z , let two actions have utilities $(1, 0)$ and $(0, 1)$. Then $V_Z = 1$ and $V_{\text{fix}} = 1/2$. Adding a third action with constant utility $3/4$ leaves $V_Z = 1$ but raises V_{fix} to $3/4$, reducing the observable advantage from $1/2$ to $1/4$. This establishes why action abundance alone is not an observability guarantee.

Model-pool closure The action set is therefore a controlled part of the estimand, not a list that can be extended without changing the question. Replacing $\mathcal{A}_1$ by $\mathcal{A}_2$ changes V_{fix} , V_Z , and V_{or} simultaneously; replacing the start model can additionally change the conditioning event S . A 32B or newly released model would thus require new generation, cost binding, multiplicity control, and a separately declared target. It could reveal more evaluator-only opportunity while leaving the recovered share ρ_Z unchanged or smaller. For the completed observability ladder, $\mathcal{A}$ and the controller family were held fixed so that the only intended intervention is the refinement $Z_1 \preceq Z_2 \preceq \dots$. Larger or newer models remain a separate scale- or generation-transfer study rather than an omitted tier of this one.

4.1 Finite-sample action support before feature sufficiency

The Bayes ordering above assumes that the conditional action values can in principle be learned. A finite FIT/DEV implementation has an earlier support requirement. Let a_{fix} be selected on FIT, let a_{alt} denote the other action, and define the fitted target

$$\Delta_j = U_j(a_{\text{alt}}) - U_j(a_{\text{fix}}) \in \{-1, 0, 1\} \quad (\text{S21})$$

for FIT stopped-history episode j .

Proposition 4 (Zero-support degeneracy of the frozen ladder). *For the ridge-plus-isotonic learner used in the LiveCodeBench ladder, suppose $\Delta_j = 0$ for every FIT episode. For every ridge penalty in the frozen positive grid, the calibrated action-advantage prediction is then zero for every input. If deviation requires $\hat{\Delta}(Z) > \tau$ with $\tau \geq 0$, the fitted policy never deviates on DEV or TEST, regardless of the feature sigma-field. Hence its held-out value equals the frozen fixed-action value, but this equality does not imply $V_Z = V_{\text{fix}}$.*

Proof. Every ridge normal equation has zero right-hand side because its response vector is zero. The fitted intercept and coefficients are therefore zero, both in each FIT fold and in the final all-FIT fit. All out-of-fold predictions are zero, and isotonic calibration against an all-zero response assigns value zero to its only block. The calibrated prediction is thus zero for any feature vector. The strict inequality $0 > \tau$ is false for every nonnegative threshold, so the action remains a_{fix} on every held-out episode. This reasoning never restricts the joint law of (Δ, Z) in the held-out population, and therefore cannot identify its Bayes value V_Z . $\square$

This proposition is implementation-specific and should not be confused with an impossibility theorem for outcome-blind signals. A future ladder should freeze a support-positivity gate before outcome opening. At minimum, FIT must contain a prespecified number of episodes with $\Delta > 0$; if the scientific claim is heterogeneous action selection rather than replacement by one dominant action, it must also contain a prespecified number with $\Delta < 0$. These are necessary support checks, not sufficient power or calibration guarantees. They are a post-audit design lesson and do not retrospectively alter the original five-layer matched-information protocol.

5 LiveCodeBench observability-ladder execution

Each LiveCodeBench query contributes three candidate blocks with distinct domain-separated deterministic seeds. A block becomes an episode only when one of at most three sequential Qwen14 history attempts first passes every public test. Private/full correctness is evaluator only. L1 includes structural and cost fields; L2 adds token confidence; L3 adds public-verifier traces and a FIT-calibrated score; and L4 adds agreement summaries over the observed pre-stop history. All tiers use the same fixed action, ridge family, isotonic calibration, nonnegative threshold grid, and cost gate.

For stopped block b of query i , write Y_{ib}^0 for stop correctness and Y_{ib}^a for second-action correctness, both visible only to the evaluator. For $a \in \mathcal{A} = \{\text{RESAMPLE}, \text{REROUTE}\}$,

$$U_{ib}(a) = \max\{Y_{ib}^0, Y_{ib}^a\}, \quad (\text{S22})$$

$$\hat{V}_r(\pi) = \frac{1}{|\mathcal{I}_r^+|} \sum_{i \in \mathcal{I}_r^+} \frac{1}{|\mathcal{B}_i|} \sum_{b \in \mathcal{B}_i} U_{ib}(\pi(Z_{ib})), \quad (\text{S23})$$

where $\mathcal{I}_r^+$ contains split- r queries with at least one natural stop. A controller makes exactly one second-action call for each such episode. Stop-only is a descriptive reference, not a legal controller action; the TEST value covers 70 eligible queries rather than all 137 TEST queries. FIT ridge, action-isotonic, and verifier-calibration fits weight stopped episodes equally. Fixed-action selection and DEV/TEST value, token, latency, and inference calculations first average within query and then weight eligible queries equally as in (S23).

The base `VISIBLE_ONLY` and `OPEN_OUTCOMES` feature artifacts leave the L3 calibrated-verifier field null. During controller execution the frozen FIT calibrator injects that field before action selection; no DEV/TEST correctness is used to fit it. After removing each OPEN row’s evaluator field, all 155 scientific stopped-history rows match the VISIBLE rows exactly. The artifacts differ in their `episodes`, `mode`, and `test_scoring_authority` fields, but both bind the same materializer source closure, `8a049d9814aa516baee7b1aeebc04454cb817418f87606b8aaeb8b34e344d3a1`.

Algorithm S1 Frozen LiveCodeBench observability ladder

```
1: Input: FIT and DEV stopped histories; frozen tiers  $Z_1 \subset \dots \subset Z_4$ ; sealed TEST; actions  $\mathcal{A}$ 
2: choose the FIT equal-query best fixed action; break an exact tie toward RESAMPLE
3: for  $k = 1, \dots, 4$  do
4:   construct cumulative  $Z_k$ ; standardize on FIT and append explicit missingness indicators
5:   set  $\Delta = U(\text{ALTERNATIVE}) - U(\text{FIXED})$  on FIT episodes
6:   for  $\lambda \in \{10^{-4}, 10^{-3}, \dots, 10^4\}$  do
7:     obtain five query-fold out-of-fold ridge predictions, fit isotonic calibration, and refit ridge on all FIT episodes
8:     for  $\tau \in \{0, .01, .02, \dots, .50\}$  do
9:       on DEV choose the alternative iff  $\hat{\Delta}_k > \tau$ ; require total-token ratio  $\leq 1.05$  and second-action-latency ratio  $\leq 1.05$ 
10:    end for
11:  end for
12:  freeze by highest DEV equal-query utility, fewest deviations, largest  $\tau$ , then largest  $\lambda$ 
13: end for
14: open TEST once; compare each frozen tier with fixed and a deterministic deviation-count-matched random policy
15: compute 20,000 PCG64 query-cluster replicates with seed 20260830404 (random comparison seed +1)
16: form one-sided studentized max- $T$  simultaneous lower bounds over L1–L4 and pass a tier only when both lower bounds are  $> 0$  and its cost gate passes
17: return the least informative passing tier, or the frozen no-tested-signal verdict
```

Table S3: Stopped-history support across the frozen LiveCodeBench splits. A “rescue” means that at least one legal second action turns an incorrect stop into full correctness.

Split	Queries	Stops	Incorrect	Any rescue	$\Delta > 0$	$\Delta < 0$	$\Delta = 0$
FIT	136	165	66	0	0	0	165
DEV	68	89	41	5	0	4	85
sealed TEST	137	155	49	5	3	2	150

TEST has 411 candidate blocks and stopped histories in 70 queries. Of the 155 stops, 106 are fully correct. Among the 49 incorrect stops, 44 are rescued by neither action, two by resampling only, three by rerouting only, and zero by both. “Any rescue” is measured relative to stop-only, whereas Δ compares reroute with the FIT-selected resample action. Thus DEV has five rescue episodes but no positive alternative-over-fixed target; four favor the fixed resample action and one is tied. Positive deviation support appears only in TEST. Nevertheless, Proposition 4 forces every frozen L1–L4 controller to make zero deviations.

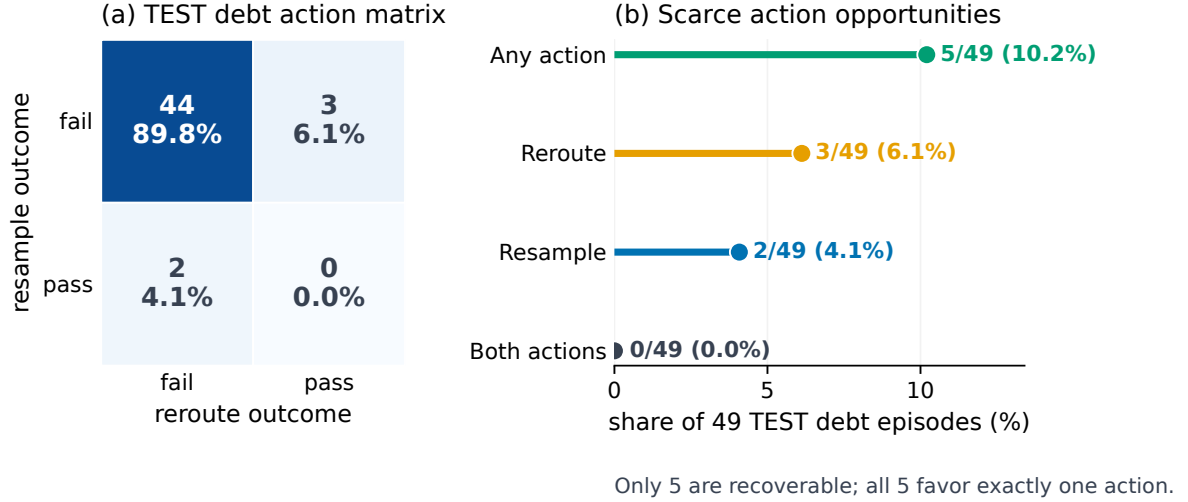


Figure S3: Action outcomes within the 49 sealed TEST stopping-debt episodes. Forty-four are rescued by neither action, three by rerouting only, two by resampling only, and zero by both actions. The five recoverable episodes show finite-bank realized rescue, but their scarcity does not establish a population rate or an observable rule for choosing the action.

Figure S3 makes the action-specific rescue count explicit without turning evaluator-only outcomes into controller inputs.

Table S4: Equal-query-weighted sealed TEST values. The reroute row is a descriptive sensitivity, not the frozen primary baseline.

Rule	Value	Gain over frozen resample	Learned deviations
Frozen fixed resample	0.595238	0.000 points	—
Fixed reroute (descriptive)	0.611905	+1.667 points	—
Evaluator-only oracle	0.621429	+2.619 points	—
L1, L2, L3, or L4 frozen controller	0.595238	0.000 points	0

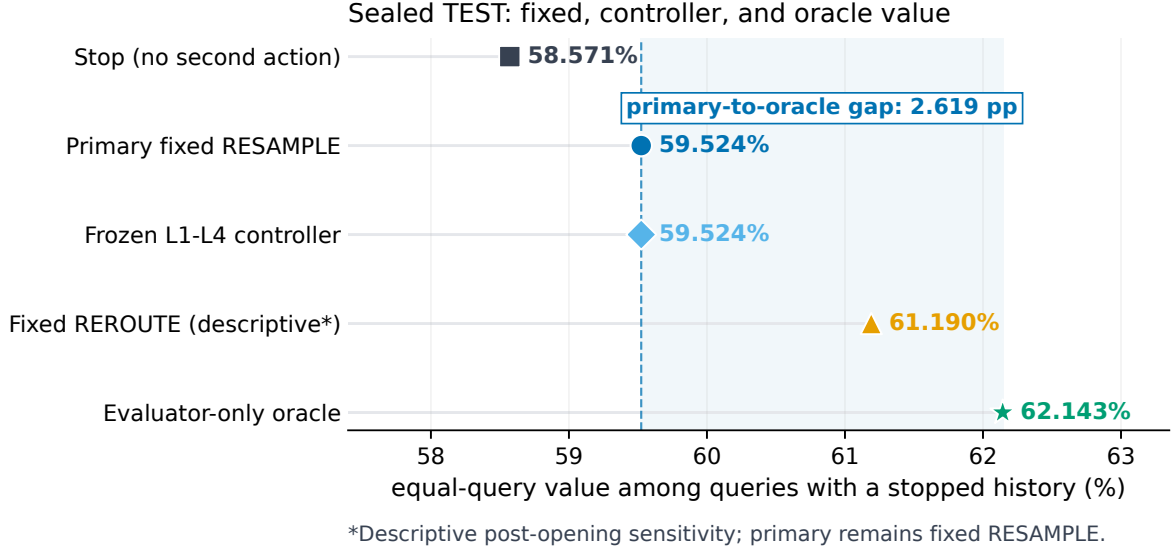


Figure S4: Sealed LiveCodeBench TEST value among the 70 queries with at least one stopped history, using equal query weight and equal stopped-episode weight within query. The frozen L1–L4 controllers overlap the preregistered fixed- resample baseline because every tier makes zero deviations. Fixed reroute is a post-opening descriptive sensitivity, and the evaluator-only oracle measures opportunity rather than achievable controller value. The shaded 2.619-point oracle–resample gap contains both fixed-action transport loss and residual evaluator-only information; it is neither a deployable gain nor a pure information gap.

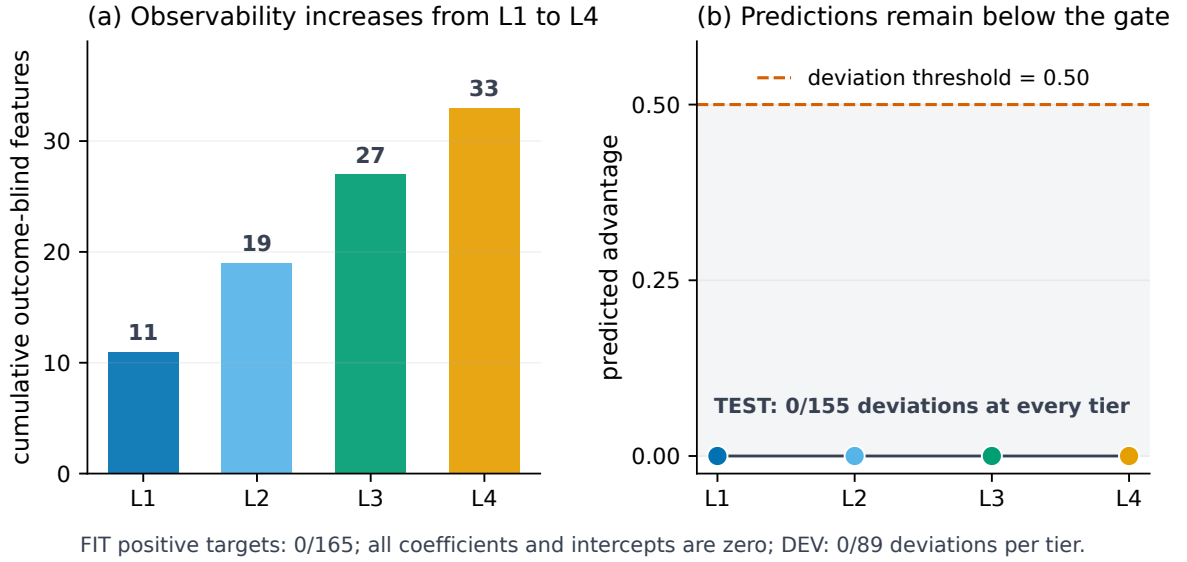


Figure S5: The frozen ladder expands from 11 to 33 cumulative scientific outcome-blind fields (22 to 66 design columns after missingness expansion), yet every fitted action-advantage prediction is zero and lies below the strict 0.50 deviation threshold. This zero-deviation path is the algebraic consequence of all 165 FIT targets being exactly zero under the frozen ridge-plus-isotonic learner; it is not evidence that the added features are generally uninformative.

Figure S5 visualizes the implementation-specific degeneracy in Proposition 4; it is not a comparison of the Bayes envelopes associated with L1–L4.

The machine conclusion is `OBSERVABILITY_LADDER_NO_TESTED_SIGNAL_BEATS_FIXED`. Its

scientific interpretation is narrower: the experiment detects finite-sample FIT support and action-ranking transport failure and cannot compare feature sufficiency. The descriptive reroute-minus-resample difference is +1.667 points; it neither establishes reroute superiority nor licenses a post-outcome baseline change. The evaluator-only residual above reroute is 0.952 points. A post hoc calculation gives a history-plus-action generated-token ratio of 0.95073 and a second-action-only latency ratio of 0.59001 for fixed reroute versus fixed resample. These descriptive ratios do not amend the utility-only FIT tie rule. The sealed schema’s historical field name `residual_information_debt_binary64_hex` stores $V_{\text{or}} - V_{\text{learned}}$; it is not an estimate of the theoretical Bayes residual $V_{\text{or}} - V_Z$ and will be renamed in a future schema revision.

6 BigCodeBench two-sided FIT support replication

6.1 Frozen design and stopping rule

BigCodeBench v0.1.4 is pinned by parquet revision, byte count, and SHA-256; the official evaluator is pinned to release v0.2.4, commit 9059fb84d1188c02edeac4995361656a2fdecbeef, and an immutable container manifest. Near-duplicate prompt components are assigned as units to one pilot, FIT, DEV, or TEST, and the 483 FIT tasks are further assigned to five folds. For each task and each of three episodes, up to three Qwen14 history attempts are generated. The first attempt passing the frozen public subset defines the stop; a stopped episode receives exactly one disjoint Qwen14 resample and one Qwen7 reroute. The resulting banks contain 895 stopped episodes in 351 queries and zero action draws in the other 132 FIT queries.

Let S_{ib} denote the already frozen public stop event and let $Y_{ib}^0, Y_{ib}^{\text{res}}, Y_{ib}^{\text{rer}}$ be one official full-suite realization for history, resample, and reroute. The two legal utilities are

$$U_{ib}(\text{res}) = Y_{ib}^0 \vee Y_{ib}^{\text{res}}, \quad (\text{S24})$$

$$U_{ib}(\text{rer}) = Y_{ib}^0 \vee Y_{ib}^{\text{rer}}, \quad (\text{S25})$$

$$\Delta_{ib} = U_{ib}(\text{rer}) - U_{ib}(\text{res}) \in \{-1, 0, +1\}. \quad (\text{S26})$$

Write

$$\mathcal{E} = \{(i, b) : S_{ib} = 1\}, \quad \mathcal{Q} = \{i : \exists b (i, b) \in \mathcal{E}\}, \quad (\text{S27})$$

$$N_s = \sum_{(i,b) \in \mathcal{E}} \mathbf{1}\{\Delta_{ib} = s\}, \quad Q_s = \sum_{i \in \mathcal{Q}} \mathbf{1}\{\exists b : \Delta_{ib} = s\}, \quad (\text{S28})$$

$$Q_{sf} = \sum_{i \in \mathcal{Q}} \mathbf{1}\{g(i) = f\} \mathbf{1}\{\exists b : \Delta_{ib} = s\}, \quad s \in \{-1, +1\}, \quad (\text{S29})$$

where $g(i) \in \{0, \dots, 4\}$ is the frozen component-level fold. For the L3 stop-confidence calibrator, define the stop-class counts

$$C_y^E = \sum_{(i,b) \in \mathcal{E}} \mathbf{1}\{Y_{ib}^0 = y\}, \quad (\text{S30})$$

$$C_y^Q = \sum_{i \in \mathcal{Q}} \mathbf{1}\{\exists b : (i, b) \in \mathcal{E}, Y_{ib}^0 = y\}, \quad y \in \{0, 1\}. \quad (\text{S31})$$

A query is sign-present if any of its stopped episodes has the sign. With $q = Q_s$ and $n = |\mathcal{Q}|$, the implementation computes the one-sided 97.5% Clopper–Pearson lower bound

$$L_s = \begin{cases} 0, & q = 0, \\ \text{Beta}^{-1}(0.025; q, n - q + 1), & q > 0, \end{cases} \quad (\text{S32})$$

by 96 deterministic bisection iterations on the exact binomial upper tail. The two bounds have Bonferroni family confidence at least 95%. The complete authorization bit is

$$\begin{aligned}
G_{\text{FIT}} = & \mathbf{1}\{|\mathcal{E}| \geq 400, |\mathcal{Q}| \geq 200\} \\
& \wedge \bigwedge_{s \in \{-1, +1\}} \mathbf{1}\{N_s \geq 25, Q_s \geq 20, \min_f Q_{sf} \geq 2, L_s > 0.01\} \\
& \wedge \bigwedge_{y \in \{0, 1\}} \mathbf{1}\{C_y^E \geq 25, C_y^Q \geq 20\}.
\end{aligned} \tag{S33}$$

These are conjunctive hard gates: total sample size, prevalence, or fold coverage cannot compensate for a failed sign count. In particular, the gate tests whether both selector labels are estimable before any selector is fit; it is not a test that resampling or rerouting has zero value.

Algorithm S2 Outcome-blind FIT gate evaluation and authorization

Require: frozen source closure, FIT generation banks, checkpoint run config, verified authority, and preregistered minima

- 1: verify byte identities, read-only modes, canonical checkpoint prefix, ordered coordinates, request continuity, and zero DEV/TEST flags
- 2: **if** any code, digest, provenance, or container invariant fails **then**
- 3: preserve all evidence and terminate; do not reinterpret the failure
- 4: **end if**
- 5: **for** each $(i, b) \in \mathcal{E}$ in frozen coordinate order **do**
- 6: accept exactly one structured official full-suite result for each of history, resample, and reroute
- 7: compute $U_{ib}(\text{res}), U_{ib}(\text{rer}), \Delta_{ib}$; record public replay and partition consistency only as diagnostics
- 8: never retry, discard, or select an outcome from a diagnostic value
- 9: **end for**
- 10: compute Eqs. (S29)–(S33) and freeze the gate input, receipt, provenance, and terminal hashes
- 11: **if** $G_{\text{FIT}} = 0$ **then**
- 12: lock L1–L4, DEV, and TEST; return the terminal support-stop verdict
- 13: **else**
- 14: authorize FIT controller fitting and DEV selection only
- 15: freeze the DEV-selected controller; permit one TEST opening only after the separate opening guard validates every binding
- 16: **end if**

6.2 Why partition equality requires common randomness

Candidate 010 originally treated the realized equality between full-suite and separately executed public/evaluator-only outcomes as fatal. That identity is valid only under deterministic tests or common evaluator randomness. Write $Y_P(c; \omega)$ and $Y_E(c; \omega)$ for public and evaluator-only outcomes of code c under execution state ω . If the full suite is exactly their union in the same realization, then

$$Y_F(c; \omega) = Y_P(c; \omega) \wedge Y_E(c; \omega). \tag{S34}$$

Separate processes instead realize $\omega_P, \omega_E, \omega_F$. Even with the same test-source partition,

$$Y_F(c; \omega_F) \neq Y_P(c; \omega_P) \wedge Y_E(c; \omega_E) \tag{S35}$$

can occur when tests, dependencies, time, process state, or generated values are stochastic. Re-running a failed coordinate until Eq. (S34) happens would accept labels conditional on a post-outcome event and bias the scientific sample.

The outcome-blind candidate-011 amendment therefore freezes S_{ib} as the stop event, takes exactly one Y_F realization per arm as the correctness label, and retains public replay and partition equality only as diagnostics. No diagnostic value triggers a retry. Code identity, task identity, source closure, container isolation, draw digests, and checkpoint continuity remain fail-closed. Candidate 010 is preserved but excluded at its intact 227-job prefix; all 895 candidate-011 jobs were rerun from zero so no selection-conditioned checkpoint was reused.

6.3 Terminal FIT result

All 895 candidate-011 checkpoints are exact canonical JSON with mode 0400 and coordinates 0–894. Each binds the checkpoint run config, HISTORY/RESAMPLE/ REROUTE draw digests, structured worker result, outcome row, and successful owned-container cleanup. The final roster contains 674 correct stops and 221 incorrect stops, the latter appearing in 118 of 351 eligible queries.

Table S5: BigCodeBench FIT stopping debt, action abundance, and signed support.

Quantity	Episodes	Distinct queries
Stopped histories	895	351
Incorrect stops	221	118
Neither action rescues	167	—
Resample only / reroute only rescues	22 / 23	19 / 19 signed support
Both actions rescue	9	—
$\Delta = +1$ / $\Delta = -1$	23 / 22	19 / 19
Hard minimum for each sign	25	20

Both signs occur in all five folds: positive/negative query counts are (5, 2), (3, 4), (5, 6), (3, 3), (3, 4). Their one-sided exact lower prevalence bounds are both 0.0329014, above 0.01. Nevertheless $\Delta = +1$ misses the episode/query minima by two/one and $\Delta = -1$ misses them by three/one. The exact verdict is `STOP_INSUFFICIENT_TWO_SIDED_FIT_SUPPORT`, with authority effect `TERMINAL_STOP_DEV_AND_TEST_MUST_REMAIN_UNOPENED`. L1–L4 are not fitted, no DEV model selection occurs, and TEST is never opened. The observability tiers are therefore unestimated rather than tied or null.

6.4 Value, concentration, cost, and diagnostic robustness

The following values are post-gate descriptions, not inputs that can revise the support verdict.

Table S6: BigCodeBench FIT values and frozen action costs. Value first averages stopped episodes within each of 351 eligible queries. Batch time comes from separate action-bank runs and is descriptive.

Rule	Equal-query value	Tokens/action	Recorded seconds/action
Stop only	0.739791	—	—
Fixed Qwen14 resample	0.771130	509.705	9.129
Fixed Qwen7 reroute	0.776353	339.710	3.899
Two-action evaluator oracle	0.799145	—	—

Reroute exceeds resample by $11/2106 = 0.522$ percentage points while using 169.995 fewer generated tokens per equal-query draw. It is also faster in the separately recorded bank

wall times. This is a FIT-sample fixed-action Pareto description, not authorization for static prompt routing and not a substitute for two-sided selector support. Episode weighting preserves the same descriptive ordering: stop, resample, reroute, and oracle values are 0.753073, 0.787709, 0.788827, and 0.813408.

Only 36 queries carry either action-advantage sign and two carry both signs in different episodes. The ten largest sign-bearing query contributions contain 14 of 23 positive and 13 of 22 negative episodes. Among eligible queries, 233 have no incorrect stop, while 50, 33, and 35 have one, two, and three incorrect stops. These distributions motivate query-level minima and query-cluster inference; they do not justify counting episodes as independent.

The rare signs occur in only 36 of 351 eligible queries; two queries contain both signs across episodes. The ten most sign-bearing queries account for 60.9% of reroute-favoring and 59.1% of resample-favoring episodes. Because these ten are selected from only 19 sign-bearing queries per direction, the shares are descriptive and are not used as evidence of exceptional concentration or a selective policy class.

Public-stop outcome replay is consistent in all 895 jobs. Separately executed suite partitions disagree in five of 2,685 completions across four jobs. This is a diagnostic of evaluator instability, not a retry gate or an estimated label-error rate. Removing those four jobs leaves 22/19 positive and 21/18 negative episodes/queries, still below the hard minima. Thus the terminal branch is robust to diagnostic exclusion, and both episode- and query-weighted descriptions prefer the same fixed action. No robustness analysis reads DEV or TEST.

6.5 Independent validation and reproduction bindings

A standard-library audit independent of the protocol package rechecks the 64-file frozen source roster, 895 contiguous checkpoint names and modes, canonical bytes, ordered coordinates, three actual bank draw digests per job, worker-result and outcome-row digests, container audits, terminal diagnostic counts, gate projections, and zero DEV/TEST flags. A second validator recomputes the exact gate receipt with the frozen candidate-011 implementation. Both pass. Publication-level identities are:

Remote 64-file source roster c7fe43b1e41b1618471d0c687ed0e487
30469e9874c096ac44cbb4885008465f

Digest of a sha256sum-format listing of the 64 read-only source files retained on the execution host. The remote tree and listing are not part of the published package.

Packaged 14-file scoring-closure roster 10d1c1a4198a4ab1da6587026f029558
1f744a8e6e524d80959f6a1e941ad8eb

Checkpoint run config b2cc49040eb334e1b60728d5d5ddc0d
7a96628fd6af4ab4a9d8d98eba44ec297

FIT scoring terminal 257f9d65cc9103c6bfaf30f4e058a6da
4c65613d52c8d0870b4e3d14e2edc72c

FIT gate input 98685d71d5ca0216be225bba3c912088d
78ee5a96d24408b810690cf93151171

FIT gate receipt a8ead2b2d87a6c25da2b756b02a1a506
1e72245d1a6bacfe22b80e5e9ed70f4e

FIT gate provenance ad0a8b35bf414fcf2cf3b8619b7af356
8ce0f9023a6c1a4ee96124ed31fd6a78

Cost projection 328aea1c2a8c585b2bc2d9516e30604d
079a6c2c4c2411a695295ad38c979ac6

Terminal analysis ab2c87b548ec506b58f90cb203fc1510
bad05d889db004a3982d013d8d2030da

The cost projection contains no source code or new correctness label. It binds every selected action draw to the terminal digest and projects only generated tokens, bytes, and query-batch elapsed time from the frozen action banks. The analysis script refuses to overwrite an existing result and recomputes every table entry above from the six canonical evidence inputs.

7 Exact arithmetic for the descriptive cost boundary

This section recomputes the main paper’s cost sensitivity from the canonical opened TEST feature artifact with the following SHA-256 identity:

e7ea286f4506e7efb1ba3327f5bdb716
6901940310cc2015052c1348b93d1c5b

It uses exact rational arithmetic, equal-query weighting over the 70 TEST queries with at least one natural stop, and no randomness. The receipt is classified POSTHOC_DESCRIPTIVE_COST_BOUNDARY; it is not a new confirmatory TEST decision.

Direct recomputation gives

$$\bar{C}_H = \frac{33259}{105}, \quad \bar{C}_A(\text{res}) = \frac{28337}{140}, \quad \bar{C}_A(\text{rer}) = \frac{6189}{35}, \quad (\text{S36})$$

$$\bar{L}_A(\text{res}) = \frac{220631209931}{70000000000} \text{ s}, \quad \bar{L}_A(\text{rer}) = \frac{65086794021}{35000000000} \text{ s}. \quad (\text{S37})$$

Therefore

$$\bar{C}_T(\text{res}) = \frac{218047}{420}, \quad \bar{C}_T(\text{rer}) = \frac{51826}{105}, \quad \frac{\bar{C}_T(\text{rer})}{\bar{C}_T(\text{res})} = \frac{207304}{218047} = 0.9507308. \quad (\text{S38})$$

The corresponding action-only token ratio is 0.8736281. The recorded action-latency ratio is

$$\frac{\bar{L}_A(\text{rer})}{\bar{L}_A(\text{res})} = \frac{130173588042}{220631209931} = 0.5900053. \quad (\text{S39})$$

Let $\lambda_T = p_T/w$, $\lambda_L = p_L/w$, and $\delta_a = q_a/w$. Substituting the exact utility contrasts into the main paper’s net-benefit definition yields

$$\delta_{\text{res}} + 202.407143\lambda_T + 3.151874\lambda_L < 0.00952381, \quad (\text{S40})$$

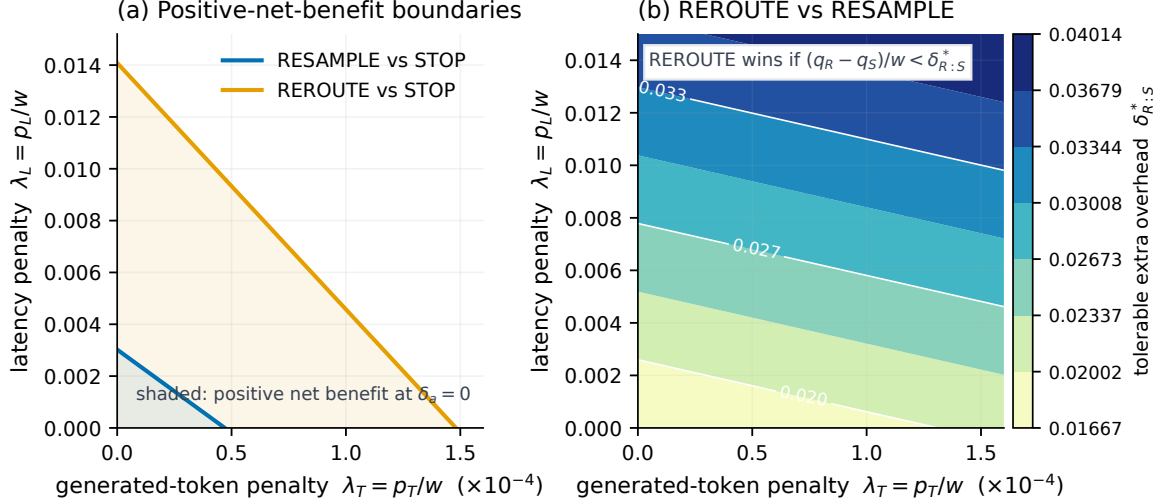
$$\delta_{\text{rer}} + 176.828571\lambda_T + 1.859623\lambda_L < 0.02619048. \quad (\text{S41})$$

The first inequality characterizes RESAMPLE versus STOP; the second characterizes REROUTE versus STOP. Their direct comparison is

$$\delta_{\text{rer}} - \delta_{\text{res}} < \frac{1}{60} + \frac{3581}{140}\lambda_T + \frac{90457621889}{70000000000}\lambda_L. \quad (\text{S42})$$

Only a sufficiently large unmeasured rerouting-specific penalty can reverse REROUTE’s advantage on the three recorded sample coordinates. Conversely, because each L1–L4 policy is action-identical to fixed RESAMPLE, $NB_{L_k} - NB_{\text{res}} = -q_k \leq 0$ for any nonnegative controller overhead.

These boundaries include generated output tokens only. They exclude input and prefill tokens, history latency, queueing, cold starts, model switching and residency, verifier and controller execution, scorer costs, dollars, and energy. Oracle cost is intentionally undefined because utility ties have no frozen physical-action or cost rule. The receipt binds every numerator, denominator, decimal rendering, source byte count, and source digest in `paper/figure_data/cost_break_even_receipt.json`.



Normalized symbolic sensitivity over an opened TEST sample; no dollar, energy, queue, or switching price is assumed.

Figure S6: Normalized post-opening cost sensitivity. Panel (a) shows the action-versus-STOP positive-net-benefit regions when unmeasured overhead is zero. Panel (b) shows the maximum additional rerouting-specific overhead that preserves REROUTE’s advantage over RESAMPLE. The axes are symbolic correctness-equivalent penalties, not observed prices. The analysis is conditional on the TEST queries with a natural stop and is descriptive, not a preregistered production Pareto claim.

8 Descriptive all-episode matched-information audit

For a fixed start-model draw (i, d) and legal second action a , write $d^+ = (d + 1) \bmod D$ and

$$U_{ida} = \max\{Y_{id}^s, Y_{id^+}^a\}. \quad (\text{S43})$$

Let a_{-i}^{fix} be the best fixed second action selected without the held-out query and let $a_{id}^* = \arg \max_a U_{ida}$. The original audit reported

$$\hat{\Delta}_{\text{rm}} = \mathbb{E}[U_{ida_{id}^*} - U_{ida_{-i}^{\text{fix}}}], \quad (\text{S44})$$

$$\hat{\Delta}_{\text{policy}} = \mathbb{E}[U_{id\hat{\pi}_{-i}(X_{id})} - U_{ida_{-i}^{\text{fix}}}], \quad (\text{S45})$$

For arbitrary two-action utilities, the realized-maximum statistic satisfies Eq. (S8); it contains no stopped-history signal and does not test conditional heterogeneity. The exchangeable reference calibrates the fold-specific fixed-action procedure under action-label symmetry.

All six probes select 1,520 episodes. The observable MBPP+ stop occurs on 1,240 episodes and the other 280 are verifier rejections. The 198 false-positive stops are an evaluator-only subset of the observable stops. The audit’s implicit event is therefore $S = \Omega$, not the MBPP+ stopping event.

The audit reuses the frozen MBPP+ raw and evaluator banks but asks a different one-step question. Every episode begins with one Qwen2.5-14B completion at draw position d . The second action either reroutes to Qwen2.5-7B or resamples Qwen2.5-14B at the fixed position $(d + 1) \bmod 10$. The primary costs are the frozen nominal parameter-scale values 70 and 140 tenths; hence the best fixed action and all cost tie-breaks are defined before fitting. No additional generation or correctness execution is required.

We evaluate three nested information regimes. NULL supplies only an intercept. STRUCTURE adds input and generated-token counts, byte and line counts, finish reason, fenced-code extraction, delimiter balance, Python parse status, expected-function coverage, and 13 AST node counts. STRUCTURE+HASH adds a normalized 256-dimensional signed hash of code token unigrams and bigrams. Raw text, token IDs, reference solutions, hidden tests, and correctness are not passed directly to the estimator. The visible projection is physically materialized by a function that accepts only the raw generation bank; a separate evaluator loader accepts only score receipts.

Five deterministic outer query folds provide held-out estimates. Within each outer training partition, four inner query folds choose $\lambda \in \{0.1, 1, 10, 100\}$. All ten positions of a query remain together. Paired percentile intervals use 10,000 deterministic query-cluster bootstrap replicates.

The post-audit exchangeable reference independently swaps the two action labels within every episode, preserves these exact outer folds, reselects each fold’s fixed action on its training queries, and recomputes the realized-maximum gap. We use 10^6 deterministic permutations with seed 20260902. This analysis was added after the adversarial audit and is reported as calibration, not as a preregistered test.

To calibrate the maximum, independently exchange the two action labels within each episode with probability $1/2$. For each exchange vector ξ , preserve the recorded query folds, reselect a_i^{fix} on the corresponding training queries, and recompute Eq. (S44). If $G(U; \mathcal{F})$ denotes that complete cross-fitted operation, the post-audit reference distribution is

$$\mathcal{N}_{\text{exch}} = \mathcal{L}_{\xi}(G(\text{swap}_{\xi} U; \mathcal{F}) \mid U, \mathcal{F}). \quad (\text{S46})$$

Algorithm S3 Descriptive all-episode matched-information audit

- 1: **Input:** frozen raw bank R , evaluator bank Y , actions A , costs c , query IDs, ridge grid Λ
 - 2: materialize legal features $X \leftarrow f(R)$ without an evaluator argument
 - 3: validate and join R and Y by query, draw, file, seed, byte length, and completion digest
 - 4: compute U_{ida} by (S43), fixed-action values, and realized-maximum values on all episodes
 - 5: **for** each deterministic outer query fold o **do**
 - 6: choose $\lambda_o \in \Lambda$ by inner query-fold validation on o^c
 - 7: fit action-utility regressions on all episodes from o^c
 - 8: select held-out actions using predictions and costs only
 - 9: evaluate selected, fixed, mixture, and oracle utilities on fold o
 - 10: **end for**
 - 11: aggregate by query and compute deterministic 10,000-replicate paired query-cluster intervals
 - 12: **Post-audit calibration:** compute the exact-fold exchangeable reference in Eq. (S46) with `paper/build.exchangeable_action_null.py`
 - 13: verify robustness and evidence invariants
 - 14: **return** descriptive gaps, exchangeable calibration, and claim boundary
-

Algorithm S4 Exact-fold exchangeable-actions reference

```
1: Input: two-action utility tensor  $U$ , recorded outer query folds  $\mathcal{F}$ , action-cost tie break
2: for  $b = 1, \dots, 10^6$  do
3:   independently swap the two action labels in every episode with probability  $1/2$ 
4:   for each recorded outer fold do
5:     reselect the fixed action on the permuted training queries
6:     score realized maximum minus that action on the held-out queries
7:   end for
8:   aggregate the held-out gap over all 1,520 episodes
9: end for
10: return the null mean, percentile interval, and observed percentile
```

The exact construction contains 96 discordant episodes. It reproduces the reported observed statistic as $41/1,520 = 2.697368$ points. Its analytic exchangeable mean is $96/(2 \times 1,520) = 3.157895$ points. With seed 20260902, one million permutations give mean 3.157851 points, standard deviation 0.448374, Monte Carlo standard error 0.000448, and 95% interval $[2.434211, 3.947368]$. The observed statistic is 0.460482 points below the null mean; 18.283% of null replicates are no larger.

The 96 discordances decompose as follows. Correct stops contribute $0 = 0 + 0$; false-positive stops contribute $21 = 14 + 7$; and verifier rejections contribute $75 = 41 + 34$, where each sum is reroute-only plus resample-only. Thus 75/96 discordances occur outside the observable stopped population.

On the descriptive restriction $Z^{14} = 1$, rerouting, resampling, and the pointwise maximum have 1,080, 1,073, and 1,087 successes. The realized-maximum gap is $7/1,240 = 0.564516$ points and the analytic exchangeable mean is $21/2,480 = 0.846774$ points.

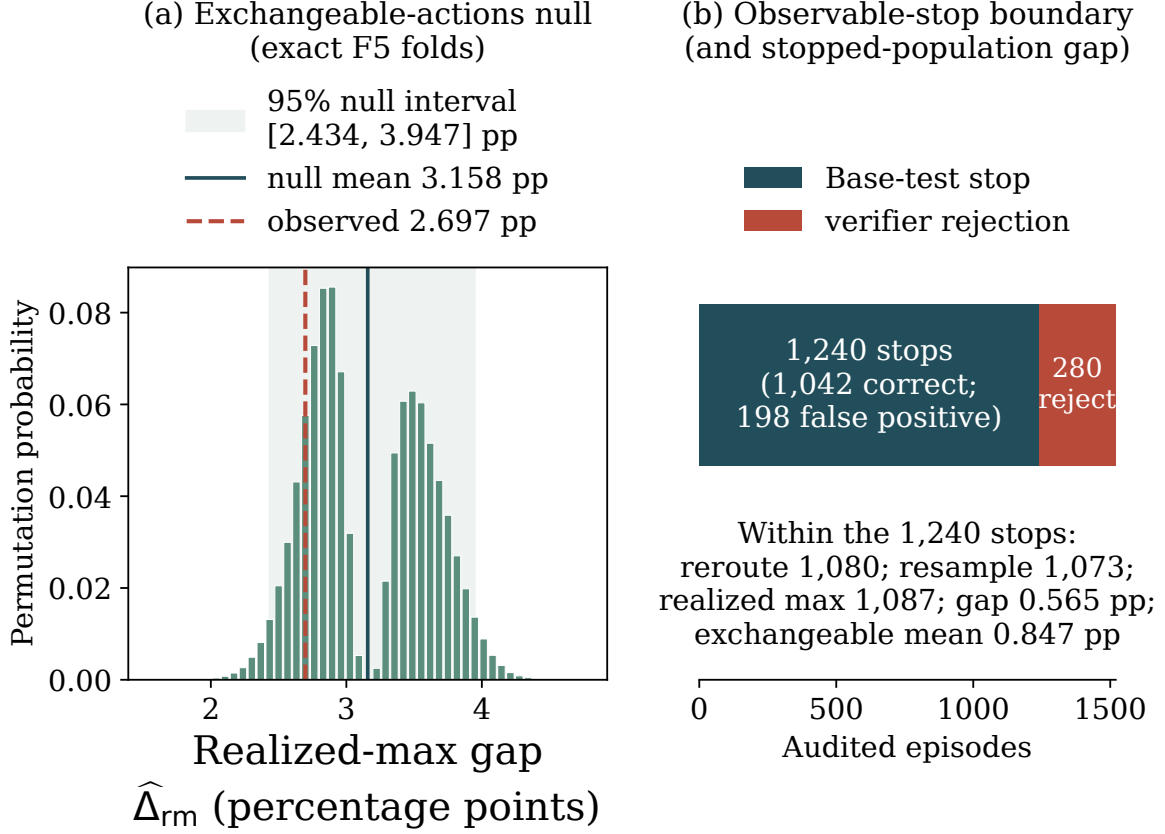


Figure S7: Exchangeable calibration and the observable-stop boundary. The reference distribution calibrates marginal label symmetry rather than conditional heterogeneity. The second panel separates observable stops from verifier rejections and reports the stopped-population diagnostic.

The structure-only controller gap remains 0.000 points $[-0.987, +1.053]$, and the structure-plus-token-hash gap remains -0.197 points $[-1.184, +0.789]$. These valid descriptive comparisons show that neither tested outcome-blind learner improves on the cross-fitted fixed action. They do not establish equality of the fixed actions, absence of conditional heterogeneity for all legal signals, or an observability bottleneck. The frozen machine-readable verdict **OBSERVABILITY_BOTTLENECK** is preserved as execution provenance; the article explicitly retracts its causal interpretation.

The post-outcome four-model equal-synthetic-cost sensitivity reported a $+4.539$ -point realized-maximum gap and negative controller gaps. It was constructed after outcome opening, has no exchangeable calibration in this study, and is not used as evidence for heterogeneity, observability, or efficiency. The four-model sensitivity added post-outcome Llama and Nemotron banks as second actions. Because it was constructed after outcome opening and used an equal synthetic cost solely for tie-breaking, it is confined to the Supplementary Material as exploratory provenance and is not used to support action heterogeneity, observability, or efficiency.

9 Deterministic cross-fitting and clustered inference

Query IDs are ordered by a domain-separated SHA-256 hash and assigned round-robin to five outer folds. Inner selection uses four query folds and a frozen ridge grid. All ten draw

positions from the same query remain in the same outer fold, inner fold, and bootstrap cluster. Action ties resolve first by lower nominal cost and then by frozen model index.

For bootstrap replicate b and sample position j , the resampled query index is the first eight bytes of

$$\text{SHA256}(\text{domain} \parallel \text{U64BE}(b) \parallel \text{U64BE}(j)) \bmod N. \quad (\text{S47})$$

Replicates are merged in increasing b . The 24-worker report is result- identical to an independent one-worker recomputation after normalizing only the recorded worker count.

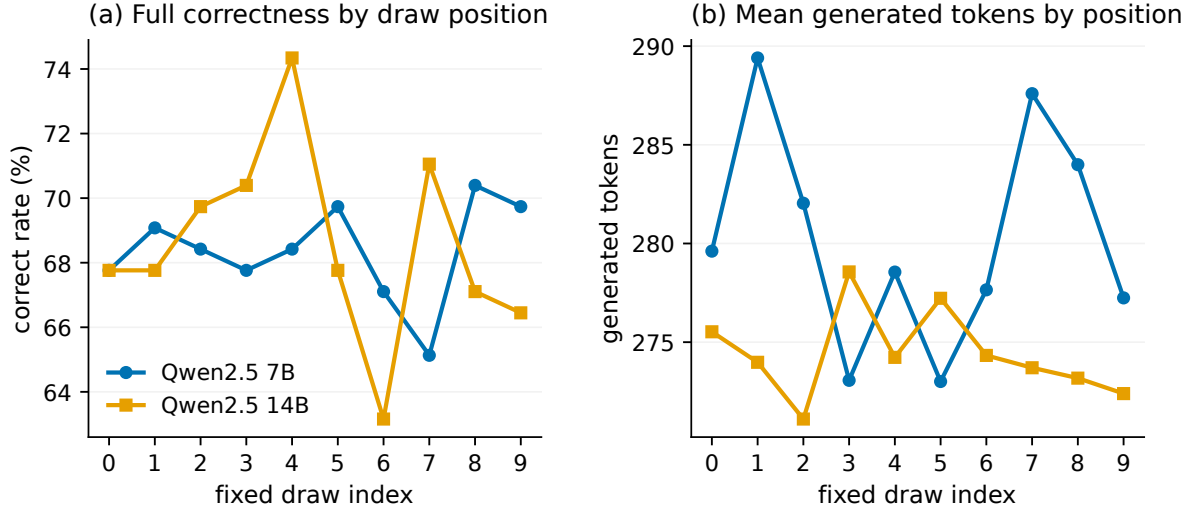


Figure S8: Correctness and generated-token diagnostics across the ten fixed draw positions. Holm correction across the four tests rejects none; non-rejection does not establish exchangeability.

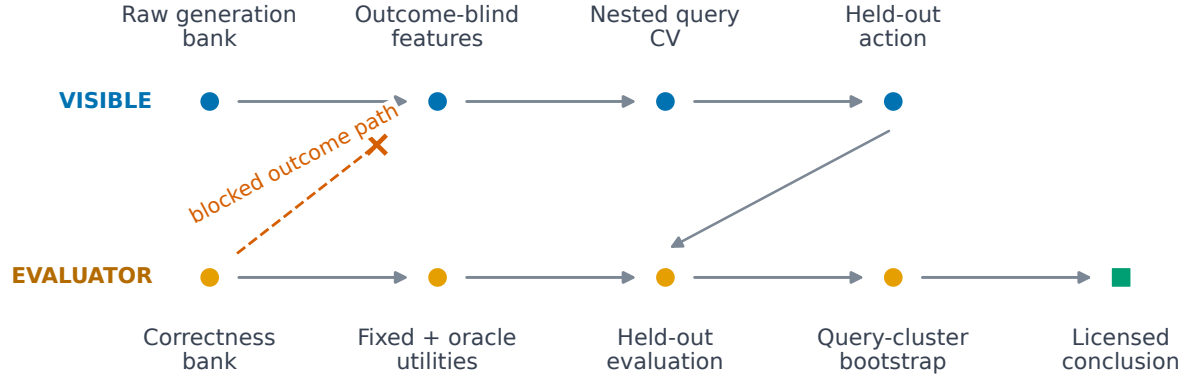


Figure S9: Matched-information audit flow. The upper lane is the only controller-visible path. The lower evaluator lane supplies training targets on training folds and held-out utility after an action has been selected, but has no path into held-out feature materialization or action selection.

10 Fail-closed self-attacks

Table S7: Final adversarial checks run before accepting the audit conclusion.

Invariant	Attack or independent check	Result
Information separation	visible materializer accepts raw generation only; selector accepts predictions and costs only	PASS
Canonical identity	strict schemas and byte-for-byte canonical JSON reproduction for all run envelopes, query files, and score banks	PASS
Completion foreign keys	query/binding, raw-file digest, draw, seed, byte length, and completion digest for 6,080 cells	PASS
Corruption response	changed completion and query-file digests rejected before analysis	PASS
Draw ledger	repeated draw, future jump, unavailable head, and over-budget action rejected	PASS
Closed form	(S6) equals exhaustive legal interleavings on all tiny banks	PASS
Parallel determinism	one-worker and 24-worker 10,000-replicate reports match after worker-field normalization	PASS
Targeted suite	known-answer STOP/simple/dynamic cases plus TEST adapter attacks	20/20

11 Cross-paper data and contribution ledger

The companion manuscript is publicly available as “How Much of the Routing Gap Is Real? Decomposing the Router-to-Oracle Gap into Reproducible Specialist Advantage and Single-Draw Label Noise” (arXiv:2607.03436); a revised version is a distinct companion manuscript. It studies coupling, scorer choice, and single-commit routing-oracle ceilings. The present article instead studies a sequential verifier-gated failure event, action- ranking support, and the conditions under which an outcome-blind controller could be evaluated. Shared historical inputs and scorer-disagreement counts are disclosed as provenance, not new contributions: the shared statistics are the historical scorer label-change counts 1,874 of 165,000 (GSM8K), 6,943 of 165,000 (MATH-500), and 5,873 of 65,340 (GPQA-Diamond) score-array entries, which also appear in the companion manuscript. Neither the older public preprint nor the unpublished revised companion is used as scientific support for the present results. A future accepted reference may identify the revised companion, but it is not needed to establish this article’s evidence or claims.

The separate arXiv preprint “Resample or Reroute? Budget-Aware Test-Time Model Selection for Large Language Models” (arXiv:2607.08665v2) is the earlier replay-based version of this research line. The present article is a substantive methodological rebuild: it does not import that preprint’s learned-router, retrospective cost-oracle, or universal Pareto claims, and it replaces them with a hidden-truth boundary, verifier-gated estimands, fresh MBPP+ evidence, the support-aware matched-information audit, and a separated LiveCodeBench ladder. The old archive remains visible only as audit/development provenance.

Table S8: Cross-paper separation and data-reuse ledger.

Asset or claim	Origin	Disposition in this article
GSM8K/MATH/GPQA response texts	shared historical generation archive	provenance inputs and shared scorer-disagreement counts; IDs, channels, and hashes disclosed
Routing-oracle decomposition	arXiv:2607.03436 companion	related-work disclosure only; not relied on, re-proved, or headlined
Earlier RoR replay and broad policy claims	arXiv:2607.08665v2	prior-version provenance; unsupported claims are not carried forward
Fresh Qwen MBPP+ banks	this study	confirmatory verifier-gated TEST evidence
Fresh Llama/Nemotron banks	post-outcome extension	model-family robustness only; not independent replication
LiveCodeBench v5+v6 ladder	this study	separated FIT/DEV/sealed-TEST support and observability diagnosis
BigCodeBench v0.1.4 FIT gate	this study	preregistered support replication; DEV and TEST remain unopened
Verifier-gated estimands	this study	headline sequence-specific mechanism contribution
Matched-information audit	this study	descriptive all-episode realized-maximum audit with exchangeable calibration
Figures and principal theorems	separate manuscripts	no figure, headline result, or principal theorem reused

12 Frozen artifact bindings

The publication-level artifacts are bound by these SHA-256 identities:

Fresh primary result a79b39ac4b40df6bd4c6469f23f820ac
a00a4a90b05cab3404b8eab2b015968

Draw-index diagnostic 4c3141bc0af60581235a422ee0560f92
88e75f8dbca7cf2942cfdcd241577d88

Model-family extension 8103ae371f624c6812445bd7072002297
fd3c5375c09451d5864a9dcf95ac14d

Matched-information audit c566ddd4069c2a7e364fab73b59545a8
2528f840c4c91480350923f1092c6858

Exchangeable-null receipt 3e617fb1fb3f019b40058d055ac626a4
c302da209d522a20e9d2532f7618fe49

Fresh final authority 4c939c360e24fb1ad5eba469e6807d46
38fef7975e43f54756a552a0d5145ac4

Qwen7 score bank 26aa52bd41f29f77669981a0ae3bcc4e
74b8b5a3dcc2091c36e7485863cae078

Qwen14 score bank 8038cb37326a72d1d3a20e5703f299cb
995ea17b868a4de10ccfa95f99d1f7a7

Llama score bank 859bd0bfeae0d32a94f9ef0c0463696
352c5faeb8e6b7af7bc0ccd359ddbafdb

Nemotron score bank 35e8f699a71529f30c6f85ee1d80e130
1b2101a1ee0bb3458f06345320dafcba

LiveCodeBench ladder controller 41bb026d42204b2435941126991ccda4
57ef477569d97d92cc5ce05237873a53

LiveCodeBench TEST-opening claim 3c64e09d72aa0e3f85691294d6f2030
a6376f78895cc18bf2ace76fcbb6944c3

LiveCodeBench TEST score config eb181e81f8c0d79d714ffb77742831cd
6a998cc48ae509ae4c2fe6200b907151

LiveCodeBench TEST score terminal 4b17d45c6fa2dfdd19e2c6b88e72674
33e0011acaffbaf932415a2b8a1556e95

LiveCodeBench visible-only features 8aac81dc5218bf0930f78abaf0d95f83
8f4a5b0512b732678995cb5dbcf5ce3f

LiveCodeBench opened features e7ea286f4506e7efb1ba3327f5bdb716
6901940310cc2015052c1348b93d1c5b

LiveCodeBench sealed result 2711bda6f4ae6abf63cffbacf7cec5bc
1ea3166bf4a09de467e7664250ee3e7d

The matched-information report binds 6,080 completion foreign keys across four raw-generation and score banks. The historical empirical manuscript and supplement reference 12 active figure PDFs: five are recorded in `figs/kbs_figure_manifest.json`, while the remaining seven are bound by the observability, cost-boundary, BigCodeBench, and exchangeable-action receipts or manifests and checked by `paper/validate_kbs_package.py`. No single 12-figure manifest is claimed. The new synthetic figures have separate paper-facing bindings described in Section 15; they do not modify this historical roster.

13 Historical replay audit boundary

The legacy 11-model replay over GSM8K, MATH-500, GPQA-Diamond, and HumanEval+ (M. Chen et al., 2021; Cobbe et al., 2021; Hendrycks et al., 2021; Lightman et al., 2024; Liu, Xia, Wang, & Zhang, 2023; Rein et al., 2024) is retained solely to document why broad allocator, static-routing, and truth-guided-verifier claims were removed. Its fixed-bank permutations are not independent datasets, parameter-count costs are not prices, and the historically inspected split is not confirmatory evidence. Observable verifiers alter trajectories, agreement can terminate on wrong answers, and same-cap comparisons can be dominated elsewhere on the measured grid. These results are subtractive provenance and do not pass or fail a headline gate.

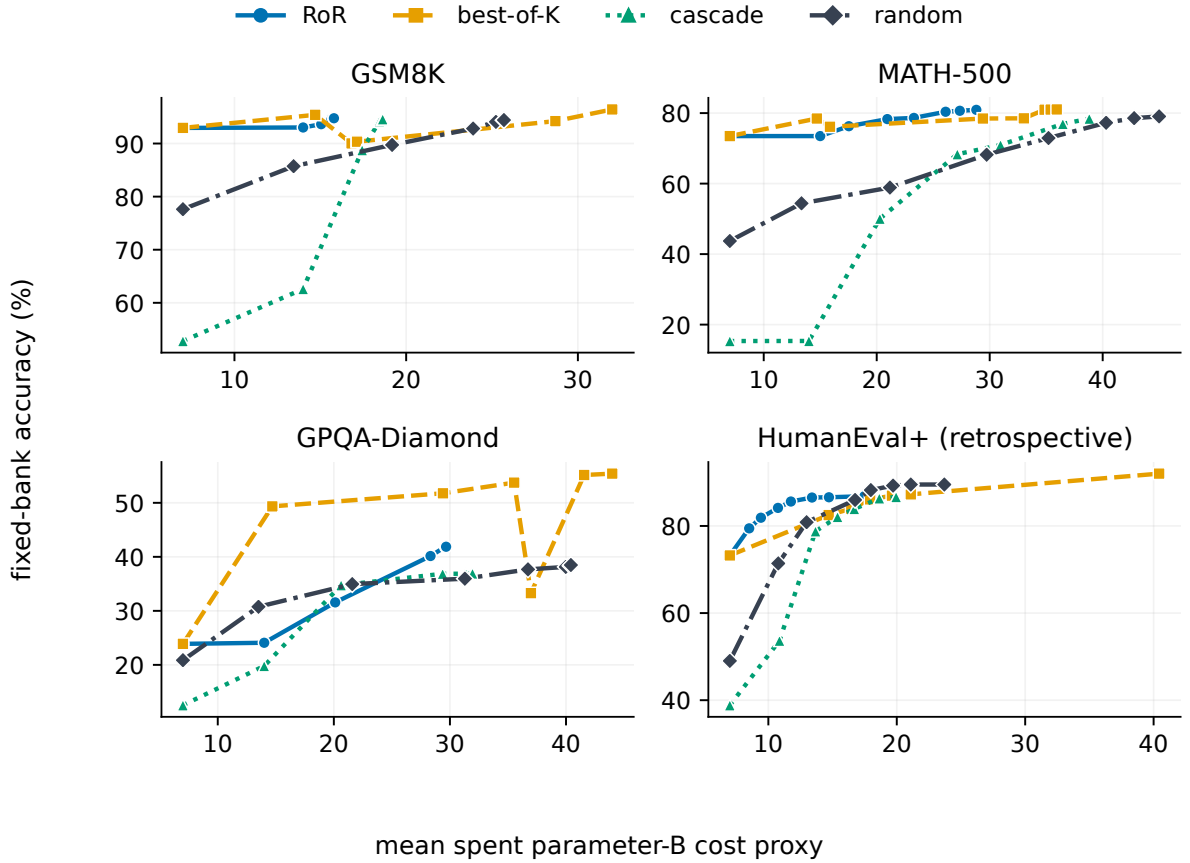


Figure S10: Policy performance under observable verifier adapters. Unlike the legacy q interpolation, these trajectories never expose held-out hidden correctness to allocation. For legibility, lines connect evaluated points in increasing realized mean cost; they are not, by themselves, Pareto frontiers. Each panel has independently scaled cost and accuracy axes, so slopes and distances must not be compared across benchmarks.

14 Interpretation checklist

For every future table, figure, or claim derived from this work, verify:

1. Which frozen response archive, split, scorer channel, and model order produced evaluator truth?
2. What information was policy visible at each decision, and what remained evaluator only?
3. Did all compared policies use the same stopping event, target population, action set, utility, and observable signal?
4. Were every query's draw positions kept together in training, cross-validation, and uncertainty estimation?
5. Is the cost axis a nominal cap, spent-call proxy, generated tokens, measured latency, or monetary price?
6. Was every pointwise realized maximum calibrated against an appropriate exchangeable reference and kept distinct from expected action heterogeneity?

7. Is the result confirmatory, cross-fitted on an opened TEST population, or post-outcome sensitivity analysis?
8. Does any new conclusion exceed the frozen artifact, model, verifier, and draw-order scope?
9. Before interpreting a null fitted controller as a feature result, did FIT contain prespecified positive and negative action-advantage support?

15 Synthetic support-gate stress test

15.1 Scope and common stopped population

This is an author-frozen synthetic support-gate stress test, not an external registration, an LLM experiment, or a validation of the historical three-gate pipeline. It transports only the numeric BigCodeBench Gate-2 minima. The generative model, query folds, learner, Gate-1 statistic, and held-out rule are new and deliberately simple. No empirical response bank supplies simulation inputs; no historical claim is relabeled and no empirical DEV/TEST split is reopened. The experiment tests an operational evidence standard, not whether all possible selectors are learnable.

Independent queries are clusters. Query i has D first-response histories indexed by $d = 0, \dots, D - 1$, one query-level post-response feature X_i , and two independent standard-normal history-level nuisance features z_{id1}, z_{id2} . Write $W_{id} = (X_i, z_{id1}, z_{id2})$. Hidden parent correctness is $Y_{0,id} \sim \text{Bernoulli}(0.6)$; the observable stop indicator S_{id} has probability 0.9 when $Y_{0,id} = 1$ and 0.6 otherwise. Therefore

$$\Pr(S = 1) = 0.78, \quad \Pr(Y_0 = 0 \mid S = 1) = \frac{4}{13}. \quad (\text{S48})$$

Only histories with $S = 1$ are action eligible. The same event defines all gates and values, and no selection depends on hidden correctness or action advantage. Queries without a stopped history receive no weight. For $n_i = \sum_d S_{id} > 0$, empirical values average the n_i stopped histories within each query and then weight eligible queries equally. X is independent of the parent/stop process. These features represent constructed post-response observations, not static prompt routing or query-identity features.

For $a \in \{\text{res}, \text{rer}\}$, an alternative has correctness Y^a , and

$$U_a = \max(Y_0, Y^a), \quad \Delta = U_{\text{rer}} - U_{\text{res}}. \quad (\text{S49})$$

U_a measures correctness availability in a pair, not the accuracy of an executable final-answer selection procedure. All quantities below condition on the common stop. The four regimes are:

Noise null. $X = -1, +1$ with equal probabilities; both alternative correctness probabilities are 0.5, with independent action coins.

Visible crossover. $X = -1, +1$ with equal probabilities. The pair $(p_{\text{res}}, p_{\text{rer}})$ is $(0.8, 0.2)$ at $X = -1$ and $(0.2, 0.8)$ at $X = +1$, with independent action coins.

Hidden crossover. An independent query-level latent $T = -1, +1$ controls those two probability pairs. The learner sees independent X , not T ; conditional on observable information both action means are 0.5.

Rare crossover. $\Pr(X = -1, 0, +1) = (0.025, 0.95, 0.025)$. At the two nonzero values use the visible-crossover probabilities and independent coins. At $X = 0$ both probabilities are 0.5 and the actions share one uniform random variable, so their outcomes are identical. This idealized construction changes action-outcome coupling as well as signal prevalence.

15.2 Population values and exact policy truth

Let $m(W) = \mathbb{E}[\Delta \mid W, S = 1]$. With all expectations taken under the same stopped population, the two-action identities are

$$V_{\text{obs}} - V_{\text{fix}} = \frac{1}{2} \{ \mathbb{E}|m(W)| - |\mathbb{E}m(W)| \}, \quad (\text{S50})$$

$$\mathbb{E} \max(U_{\text{res}}, U_{\text{rer}}) - V_{\text{fix}} = \frac{1}{2} \{ \mathbb{E}|\Delta| - |\mathbb{E}\Delta| \}. \quad (\text{S51})$$

The first expression maximizes expected utility using observable information; the second maximizes realized outcomes with hindsight. They are elementary identities, not interchangeable estimands. Every regime has $V_{\text{fix}} = 11/13$. In the order Noise, Visible, Hidden, Rare, observable optimal gains are $0, (4/13)(0.3), 0, (4/13)(0.015)$; the population realized-max gaps are $(4/13)(0.25), (4/13)(0.34), (4/13)(0.34), (4/13)(0.05)(0.34)$. Figure S13 displays these analytic values.

For frozen coefficients (b, b_x, b_1, b_2) and threshold τ , exact population evaluation integrates the Gaussian nuisances:

$$\Pr(\text{deviate} \mid X = x) = \Phi \left(\frac{b + b_x x - \tau}{\sqrt{b_1^2 + b_2^2}} \right). \quad (\text{S52})$$

At zero nuisance variance, use the strict deterministic threshold. Integrate over the declared X distribution, hidden T where applicable, and stopped parent correctness. No additional oracle sample or truth-dependent selection is used. The reported proposed and deployed gains are relative to the population best fixed action, separately from empirical FIT-reference gains.

15.3 Frozen learner, gates, and simulation grid

Method A selects the best fixed action on FIT by equal-query utility, with ties choosing resample. Method B fits a ridge selector without early screening. Method C applies synthetic Gate 1 and transported Gate 2 before using the same selector; on early failure it uses A. The synthetic rules are:

1. Gate 1 passes if either action has a strictly positive one-sided 97.5% paired-query Student- t lower bound for $U_a - Y_0$ (two-action Bonferroni allocation). All four regimes contain recoverable debt, so this is only a positive control, not a measurement of Gate-1 type-I error.
2. Gate 2 requires at least 200 stopped FIT queries and 400 episodes. Each sign of Δ must occur on at least 25 episodes and 20 queries, including at least two queries in each of five folds. Each sign's exact one-sided 97.5% Clopper-Pearson lower bound on query prevalence must be strictly greater than 0.01. Each parent-correctness class must cover at least 25 episodes and 20 queries. Folds are assigned by $i \bmod 5$ before outcomes; this is not the historical component-hash fold construction. These counts do not establish observable conditional crossover.
3. For both B and eligible C, the selected policy must have strictly positive one-sided 97.5% paired-query Student- t lower bounds against *each* fixed action on one TEST bank. The two comparisons receive a Bonferroni allocation; policy, reference, model, and threshold are fixed before TEST. Failed promotion falls back to A. These bounds are approximate, not universal finite-sample guarantees.

The target of ridge regression is the per-history utility gain of deviating from A. For design vector $v_{id} = (1, X_i, z_{id1}, z_{id2})$ and target $g_{id} = U_{\text{alternative}, id} - U_{A, id}$, fit

$$\hat{\beta}_\lambda = \arg \min_{\beta} \sum_{i: n_i > 0} \frac{1}{n_i} \sum_{d: S_{id}=1} (g_{id} - v_{id}^\top \beta)^2 + \lambda \sum_{k=1}^3 \beta_k^2. \quad (\text{S53})$$

The intercept is unpenalized. The fixed grids are

$$\lambda \in \{0.01, 1, 100\}, \quad \tau \in \{0, 0.01, 0.025, 0.05, 0.1, 0.2, 0.5\}.$$

Deviation occurs exactly when $v^\top \hat{\beta}_\lambda > \tau$. DEV selects the pair with highest equal-query utility; ties favor fewer deviations, then larger τ , then larger λ . No isotonic calibration or historical L1–L4 controller is used. Early-blocked C has no separate learning or selection computation. Because B uses the same simulation harness, C’s possible reduction in labels is a stand-alone logical count, not measured runtime or API savings.

The primary grid uses four regimes and $D = 3$, with raw FIT query counts 200, 500, 1,000, or 2,000: 500 independent repetitions per cell, totaling 8,000 jobs. DEV uses 500×3 and TEST 2000×3 raw query histories. The 200-query FIT cell is a deliberate boundary cell, not 200 guaranteed stopped queries. The secondary rare-regime grid fixes 6,000 raw FIT history attempts across $(Q, D) = (200, 30), (500, 12), (1000, 6), (2000, 3)$, adding 2,000 jobs. Within each secondary repetition, parent and stop vectors are shared across allocations, yielding exactly equal realized stopped-history counts; feature groups and alternative noise use distinct allocation streams. Thus this comparison does not isolate query count from every other aspect of grouping. DEV and TEST sizes are unchanged.

Randomness uses PCG64 with SeedSequence master seed 2026090901, formal namespace 101, and engineering namespace 901. FIT, DEV, TEST, and repetitions have disjoint streams. The 16-job timing pilot disclosed only timing and validation metadata; source and configuration were author-frozen before that pilot. No grid, threshold, hypothesis, or sample size was selected using formal outcomes. All 10,000 planned jobs are retained, including refusals.

15.4 Complete cell results and interpretation

Tables S9–S11 report every cell. Binary rates use 95% Wilson Monte Carlo intervals; continuous means and paired C–B differences use Student- t intervals across the 500 repetitions. These describe simulation variability, not uncertainty about any empirical benchmark. Missed observable value is the regime’s analytic optimum minus deployed gain. Gate 1 passes in every run. In the two observable nulls, B promotes at most once per cell, while the support gate passes in every run when $Q \geq 500$. Hence abundant two-sided realized discordance need not imply observable advantage. C promotions are a subset of B promotions by construction, not a proof of improved calibration.

For rare crossover at $Q = 1000, D = 3$, B promotes in 480/500 repetitions (96.0%; Wilson 95% interval 93.90–97.40%), C in 0/500 (0–0.76%), and B’s true deployed gain is 0.44003 pp [0.43206, 0.44800]. The paired C–B difference is -0.44003 pp $[-0.44800, -0.43206]$. The signal is exactly linear, and the neutral 95% of queries share action coins, suppressing comparison noise. The main text’s “about one repetition in eight” independent-coin statement is only a crude normal-tail sensitivity calculation: using approximate gain 0.4615 pp and standard error 0.56 pp gives $1 - \Phi(1.96 - 0.4615/0.56) \approx 0.13$ for one comparison. It is not an additional experiment, an exact joint two-comparison promotion probability, or a replacement for the observed coupled-coin result.

In visible crossover at $Q = 200$, B promotes in 500/500 and C in 52/500; 448 repetitions fail the stopped-query minimum. For rare crossover at $Q = 200$, that boundary fails in 437/500, as marked in the main figure. At equal 6,000-attempt budgets, B promotes in 484, 483, 480, and 495 runs and C in 0, 1, 25, and 0 runs in increasing- Q order. The non-monotone C pattern reflects multiple sign, query, fold, and prevalence requirements; it is not a general learning curve. Figures S11 and S12 show the complete primary and matched-budget comparisons. The synthetic counterexample does not establish that BigCodeBench could have supported a learned router. Gate failure is a refusal under this declared protocol, not evidence that learning is impossible.

Table S9: Screening and promotion in every synthetic cell. Entries are percentages with Wilson 95% Monte Carlo intervals; 500 repetitions per row. B has no early screening; C applies both early gates. Gate 1 passes in every repetition, so early-pass rates equal Gate-2 rates. Multiply a rate by five to recover its count out of 500. Noise and Hidden promotions are false promotions.

Regime	Q	D	Gate 2 (%)	B promotion (%)	C promotion (%)
Primary grid: 3 histories per raw FIT query					
Noise	200	3	9.00 [6.79, 11.83]	0.20 [0.04, 1.12]	0.00 [0.00, 0.76]
Noise	500	3	100.00 [99.24, 100.00]	0.20 [0.04, 1.12]	0.20 [0.04, 1.12]
Noise	1,000	3	100.00 [99.24, 100.00]	0.00 [0.00, 0.76]	0.00 [0.00, 0.76]
Noise	2,000	3	100.00 [99.24, 100.00]	0.00 [0.00, 0.76]	0.00 [0.00, 0.76]
Visible	200	3	10.40 [8.02, 13.38]	100.00 [99.24, 100.00]	10.40 [8.02, 13.38]
Visible	500	3	100.00 [99.24, 100.00]	100.00 [99.24, 100.00]	100.00 [99.24, 100.00]
Visible	1,000	3	100.00 [99.24, 100.00]	100.00 [99.24, 100.00]	100.00 [99.24, 100.00]
Visible	2,000	3	100.00 [99.24, 100.00]	100.00 [99.24, 100.00]	100.00 [99.24, 100.00]
Hidden	200	3	9.80 [7.49, 12.72]	0.20 [0.04, 1.12]	0.00 [0.00, 0.76]
Hidden	500	3	100.00 [99.24, 100.00]	0.00 [0.00, 0.76]	0.00 [0.00, 0.76]
Hidden	1,000	3	100.00 [99.24, 100.00]	0.00 [0.00, 0.76]	0.00 [0.00, 0.76]
Hidden	2,000	3	100.00 [99.24, 100.00]	0.00 [0.00, 0.76]	0.00 [0.00, 0.76]
Rare	200	3	0.00 [0.00, 0.76]	92.60 [89.97, 94.58]	0.00 [0.00, 0.76]
Rare	500	3	0.00 [0.00, 0.76]	96.00 [93.90, 97.40]	0.00 [0.00, 0.76]
Rare	1,000	3	0.00 [0.00, 0.76]	96.00 [93.90, 97.40]	0.00 [0.00, 0.76]
Rare	2,000	3	0.20 [0.04, 1.12]	96.40 [94.38, 97.71]	0.20 [0.04, 1.12]
Matched budget: 6,000 raw FIT history attempts					
Rare	200	30	0.00 [0.00, 0.76]	96.80 [94.87, 98.02]	0.00 [0.00, 0.76]
Rare	500	12	0.20 [0.04, 1.12]	96.60 [94.62, 97.87]	0.20 [0.04, 1.12]
Rare	1,000	6	5.00 [3.41, 7.28]	96.00 [93.90, 97.40]	5.00 [3.41, 7.28]
Rare	2,000	3	0.00 [0.00, 0.76]	99.00 [97.68, 99.57]	0.00 [0.00, 0.76]

Table S10: True deployed gains over the population best fixed action in every synthetic cell, in percentage points. Brackets are 95% Monte Carlo Student- t intervals across 500 repetitions. The C–B column is paired. Intervals are not truncated at zero; neither benchmark uncertainty nor across-cell superiority is inferred from them.

Regime	Q	D	B gain (pp)	C gain (pp)	C–B gain (pp)
Primary grid: 3 histories per raw FIT query					
Noise	200	3	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Noise	500	3	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Noise	1,000	3	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Noise	2,000	3	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Visible	200	3	9.2027 [9.1985, 9.2070]	0.9578 [0.7105, 1.2051]	-8.2449 [-8.4920, -7.9978]
Visible	500	3	9.2193 [9.2164, 9.2222]	9.2193 [9.2164, 9.2222]	0.0000 [0.0000, 0.0000]
Visible	1,000	3	9.2284 [9.2276, 9.2293]	9.2284 [9.2276, 9.2293]	0.0000 [0.0000, 0.0000]
Visible	2,000	3	9.2307 [9.2307, 9.2308]	9.2307 [9.2307, 9.2308]	0.0000 [0.0000, 0.0000]
Hidden	200	3	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Hidden	500	3	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Hidden	1,000	3	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Hidden	2,000	3	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
Rare	200	3	0.4194 [0.4087, 0.4301]	0.0000 [0.0000, 0.0000]	-0.4194 [-0.4301, -0.4087]
Rare	500	3	0.4358 [0.4277, 0.4439]	0.0000 [0.0000, 0.0000]	-0.4358 [-0.4439, -0.4277]
Rare	1,000	3	0.4400 [0.4321, 0.4480]	0.0000 [0.0000, 0.0000]	-0.4400 [-0.4480, -0.4321]
Rare	2,000	3	0.4419 [0.4343, 0.4495]	0.0009 [-0.0009, 0.0027]	-0.4410 [-0.4488, -0.4332]
Matched budget: 6,000 raw FIT history attempts					
Rare	200	30	0.4426 [0.4354, 0.4498]	0.0000 [0.0000, 0.0000]	-0.4426 [-0.4498, -0.4354]
Rare	500	12	0.4411 [0.4336, 0.4485]	0.0009 [-0.0009, 0.0027]	-0.4401 [-0.4478, -0.4325]
Rare	1,000	6	0.4397 [0.4317, 0.4478]	0.0230 [0.0142, 0.0318]	-0.4168 [-0.4284, -0.4051]
Rare	2,000	3	0.4537 [0.4494, 0.4581]	0.0000 [0.0000, 0.0000]	-0.4537 [-0.4581, -0.4494]

Table S11: Additional synthetic diagnostics for all cells: exact population gain of the proposed policy before promotion, the empirical TEST realized-max gap, and the logical fraction of labels C could avoid if run alone. All entries include 95% Monte Carlo Student- t intervals. Logical savings are not measured CPU or API savings. Units are pp for gains and percent for labels.

Regime	Q	D	Proposed gain (pp)	Realized max (pp)	Labels avoided (%)
Primary grid: 3 histories per raw FIT query					
			0.0000	7.4292	84.26
Noise	200	3	[0.0000, 0.0000]	[7.3997, 7.4586]	[81.93, 86.59]
			0.0000	7.4363	0.00
Noise	500	3	[0.0000, 0.0000]	[7.4073, 7.4653]	[0.00, 0.00]
			0.0000	7.4443	0.00
Noise	1,000	3	[0.0000, 0.0000]	[7.4144, 7.4742]	[0.00, 0.00]
			0.0000	7.4401	0.00
Noise	2,000	3	[0.0000, 0.0000]	[7.4123, 7.4680]	[0.00, 0.00]
			9.2027	10.1464	82.98
Visible	200	3	[9.1985, 9.2070]	[10.1119, 10.1809]	[80.49, 85.47]
			9.2193	10.1747	0.00
Visible	500	3	[9.2164, 9.2222]	[10.1415, 10.2080]	[0.00, 0.00]
			9.2284	10.1590	0.00
Visible	1,000	3	[9.2276, 9.2293]	[10.1256, 10.1925]	[0.00, 0.00]
			9.2307	10.1380	0.00
Visible	2,000	3	[9.2307, 9.2308]	[10.1034, 10.1726]	[0.00, 0.00]
			0.0000	10.1448	83.53
Hidden	200	3	[0.0000, 0.0000]	[10.1100, 10.1796]	[81.11, 85.95]
			0.0000	10.1590	0.00
Hidden	500	3	[0.0000, 0.0000]	[10.1249, 10.1931]	[0.00, 0.00]
			0.0000	10.1666	0.00
Hidden	1,000	3	[0.0000, 0.0000]	[10.1314, 10.2018]	[0.00, 0.00]
			0.0000	10.1338	0.00
Hidden	2,000	3	[0.0000, 0.0000]	[10.1013, 10.1663]	[0.00, 0.00]
			0.4289	0.4577	92.60
Rare	200	3	[0.4200, 0.4377]	[0.4493, 0.4661]	[92.59, 92.61]
			0.4460	0.4519	83.35
Rare	500	3	[0.4408, 0.4511]	[0.4437, 0.4601]	[83.33, 83.37]
			0.4465	0.4517	71.43
Rare	1,000	3	[0.4404, 0.4525]	[0.4432, 0.4602]	[71.41, 71.45]
			0.4493	0.4472	55.44
Rare	2,000	3	[0.4441, 0.4546]	[0.4392, 0.4552]	[55.22, 55.66]
Matched budget: 6,000 raw FIT history attempts					
			0.4516	0.4554	55.55
Rare	200	30	[0.4471, 0.4560]	[0.4469, 0.4639]	[55.53, 55.57]
			0.4504	0.4547	55.44
Rare	500	12	[0.4457, 0.4552]	[0.4465, 0.4628]	[55.22, 55.66]
			0.4470	0.4438	52.77
Rare	1,000	6	[0.4409, 0.4530]	[0.4351, 0.4524]	[51.71, 53.84]
			0.4558	0.4517	55.56
Rare	2,000	3	[0.4526, 0.4590]	[0.4439, 0.4594]	[55.54, 55.58]

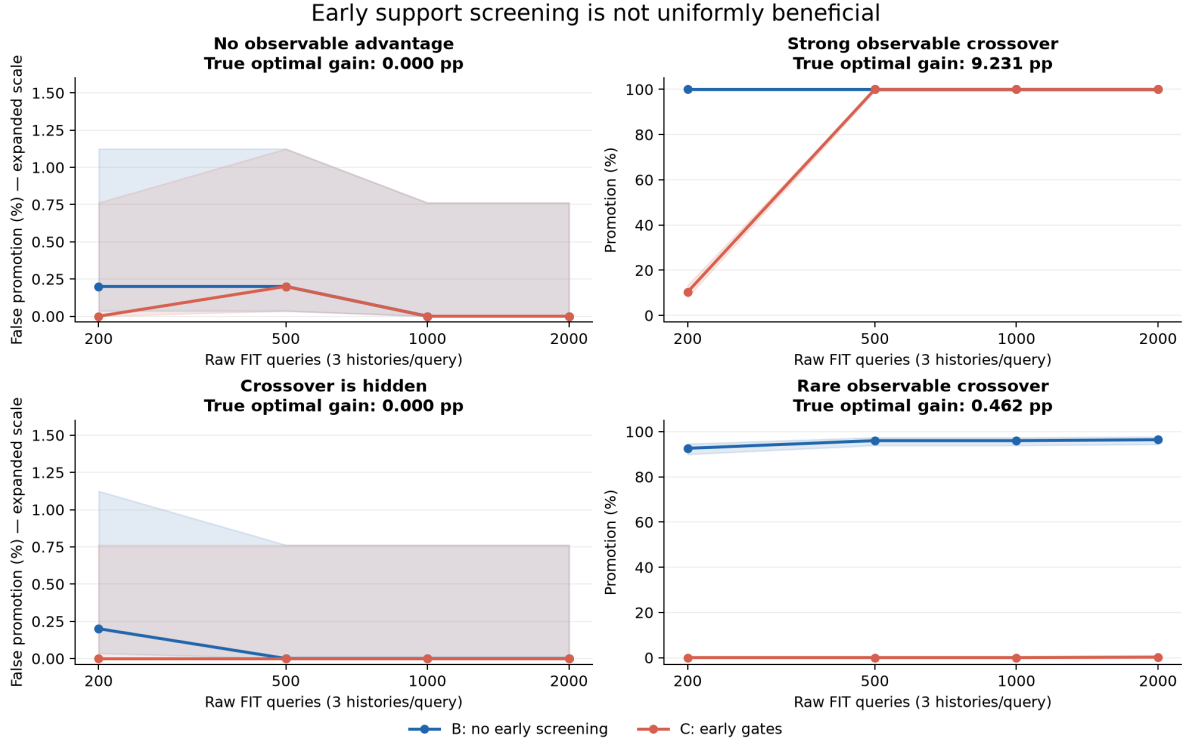


Figure S11: Synthetic support-gate stress test: complete primary grid under the author-frozen protocol, with no LLM data. Each point represents 500 repetitions; bands are Wilson 95% Monte Carlo intervals. The two null panels have expanded vertical scales and must not be visually compared with the positive-regime scales. Raw FIT counts are not stopped-query counts. The rare regime changes both signal prevalence and action-coin coupling. This figure is a byte-identical copy of the first author-report-v2 figure.

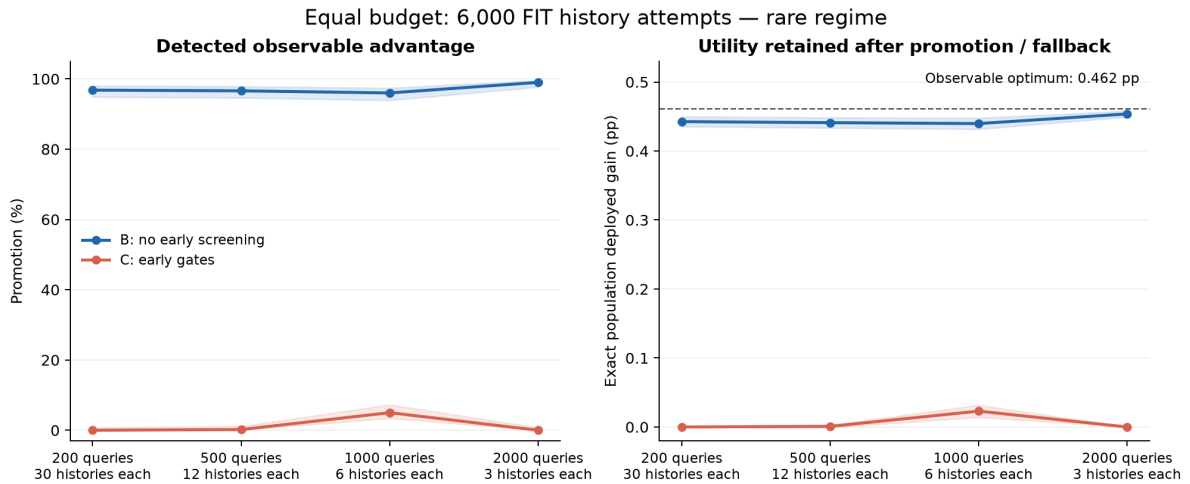


Figure S12: Synthetic rare-regime matched-budget comparison under the author-frozen protocol; no LLM data. All allocations use 6,000 raw FIT history attempts, and all 500 paired repetitions have identical realized stopped-history counts across allocations. Bands are 95% Monte Carlo intervals: Wilson for promotion rates, Student- t for mean true deployed gain. Query grouping, features, and alternative-noise streams also change, so the comparison does not isolate raw query count alone. Logical label savings are not measured compute savings. Byte-identical second author-report-v2 figure.

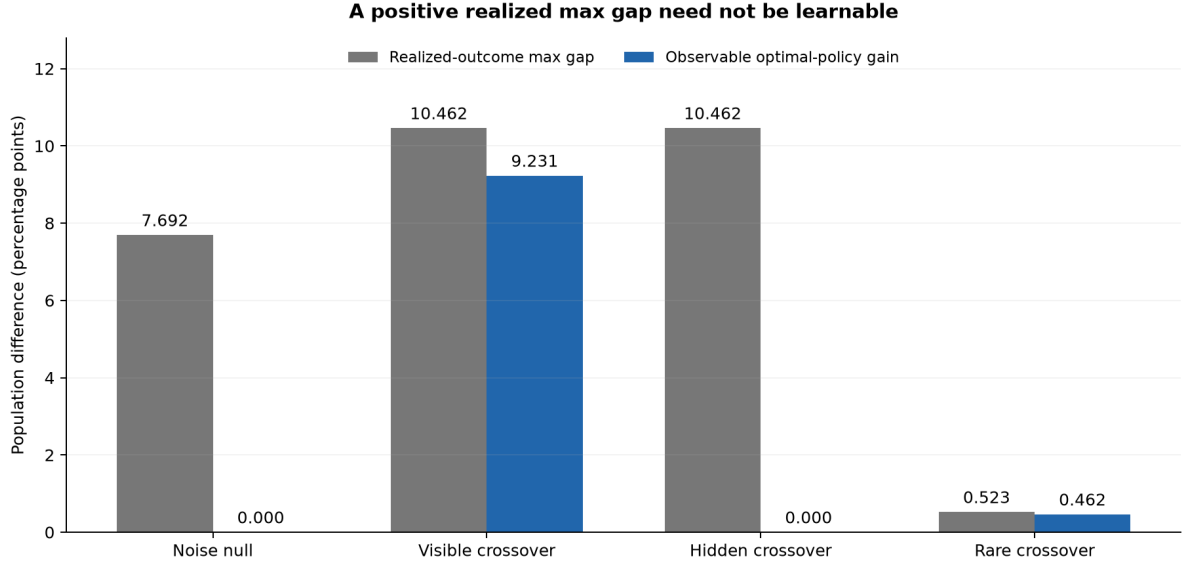


Figure S13: Analytic population values for the synthetic support-gate stress test; author-frozen protocol, no LLM data. Gray is the realized-outcome maximum gap in Eq. (S51); blue is the observable optimal-policy gain in Eq. (S50). These are known generative-model values, not finite-sample estimates, and have no Monte Carlo error bars. Byte-identical third author-report-v2 figure.

15.5 Execution and reproducibility bindings

The CPU-only run used a single Linux x86_64 host already available to the author, with at most 16 worker processes, a 16-CPU quota, single-threaded BLAS, and no GPU models or new network data. The bound environment used Python 3.12.3, NumPy 2.3.5, and SciPy 1.18.0. The formal attempt took 32.086 seconds wall-clock and 41.039 summed worker CPU-seconds; GPU time was zero. These are execution costs of this small constructed study, not estimates for new LLM evidence. Completed job records, source, configuration, and the terminal roster are preserved; the frozen study directory is not edited by manuscript production.

The immutable source closure, terminal, and all-cell summary are bound by:

Source closure 91de91c753b0893c3be06c16da81a6396
cc776b362c2fed2138e48d6db46aee3

Terminal 8387dfa0809838d135c774958f961513
49b96b471dea539b98b2b60b33f81c0e

Summary 20b1060b93676fb61e98aa064248d8ca
93661a129643523f400120324d322161

The retained directory name `mlj_gate_validation.v1/` is an immutable artifact identifier, not a scientific description of this study. Its `vm-return/source/PROTOCOL.md` and `source/config.json` specify the design; `summary.json` contains all 20 cells, intervals, per-criterion refusal counts, and additional diagnostics. The main two-panel figure reads only that summary. The three supplementary figures are copied without modification from `author-report-v2/`. Paper-facing manifests in `paper_mlj/audit/synthetic_stress_edit_20260909/` bind the summary, rendered main figure, full-cell tables, and copied figures separately from all historical empirical artifacts.

The main article’s reading map and integrated LiveCodeBench and synthetic figures were redrawn for presentation from existing manuscript statements and bound aggregate receipts, without new experiments. The original figures remain in this supplement or its preserved source assets. The new source and rendered figure bindings are recorded separately in `paper_ml_j/audit/visual_story_upgrade_20260912/`; they do not replace the historical figure manifests or frozen scientific authorities.

16 Additional execution and related-work details

These details accompany the condensed main text. They preserve the opening history and the post-outcome status of the secondary model-family extension; no experiment, scoring run, or outcome analysis was repeated for this move.

16.1 MBPP+ execution and opening chronology

Before the first sealed evaluation, the generated TEST banks, scoring inputs, scorer, runner, loader, and the then-current evaluator candidate were frozen. The first publisher execution completed both sealed scoring replays and computed the aggregate outcome using evaluator SHA `4ea5053c6c33f03483f7cbcc8872c8b6b851fd2946dd77c6d8bc3345c2d2005c`; however, strict JCS validation rejected publication because `_fraction_record` represented its `decimal` field as a Python float. Thus no canonical result artifact was published, although the sealed outcome had already been computed. After this opening, the evaluator was changed to SHA `e1a41ceb2427726adffd3ff7e62c2b27d5aae8718ba4527e905575c9d346bca2`. The change was limited to representing that field as a round-trip decimal string, and the `_query_estimands` output was byte-identical. The original authority record is retained under `invalidated/`; the fixed publisher was rerun and the canonical artifacts were frozen. Accordingly, the generation and scoring inputs were frozen before outcome opening, but the final evaluator source closure was not.

The sealed TEST opening was not authorized by an independent external reviewer. Checkpoint F2-056 requested an external Claude review and named an expected review artifact, but that review was never delivered. After the raw TEST banks were frozen, the authors issued local checkpoint F2-057 and used it to authorize the fixed scoring and opening procedure. Checkpoint F2-057 was amended after the first invalidated sealed computation to record the corrected evaluator digest; the version retained on disk therefore postdates that computation. We therefore classify the opening as self-authorized rather than independently reviewed. Its scope was limited to the two preregistered Qwen TEST banks, CPU-only scoring, the fixed zero-argument publisher, and one sealed opening; it did not authorize a 32B model, additional benchmark outcomes, pre-opening policy fitting, or static prompt routing.

The final authority re-executed both raw banks in fresh containers and constructed evaluator input only from those new receipts; stored child evidence served as provenance but not as final truth. All result files are canonical JSON, content-addressed, new-only, and frozen after publication. Recorded generation-token counts include the terminal stop token; no token-based cost enters either estimand.

16.2 Post-outcome model-family extension

After the prespecified Qwen14-to-Qwen7 outcome was opened, we froze a separate two-arm robustness protocol. It reuses the same 152 TEST queries and the same frozen Qwen2.5-14B first-stage score bank, but independently generates and rescores ten draws per query from Llama-3.1-8B-Instruct and NVIDIA-Nemotron-Nano-9B-v2 (Grattafiori et al., 2024; NVIDIA, 2025). For alternative model m , let F_i be the number of Qwen14 draws among ten that pass

Base but fail Plus and let S_{im} be the number of alternative draws that pass both. The extension estimator is

$$\hat{\Delta}_m = \frac{1}{N} \sum_{i=1}^N \frac{F_i}{10} \frac{S_{im}}{10}. \quad (\text{S54})$$

It is the same independent-pairing recovery estimand as the primary mechanism, now evaluated with two non-Qwen alternatives. We retain the fixed offset-1 pairing sensitivity. Each arm receives a descriptive 95% query-cluster bootstrap interval; the familywise decision requires both two-sided 97.5% Bonferroni interval lower bounds to exceed zero over 10,000 deterministic replicates.

The alternative banks use pinned Hub revisions and strict snapshot manifests; they were generated sequentially on one RTX 4090 (TP=1), fully unloaded between models, with Nemotron under its documented `/no_think` directive and a float32 Mamba state-space cache, and rescored in the same network-disabled EvalPlus image, while the Qwen14 bank retains its frozen TP=2 contract. The extension therefore tests model-family robustness of the finite-bank mechanism, not hardware-configuration invariance or an independent repetition of the TEST split.

Table S12: Post-outcome model-family robustness extension over the same 152 MBPP+ TEST query clusters and frozen Qwen2.5-14B first-stage bank. Recovery effects are absolute percentage points. The 97.5% intervals implement the prespecified Bonferroni familywise rule for the two extension arms.

Alternative model	Full-correct	Recovery	95% interval	97.5% interval
Llama-3.1-8B-Instruct	61.18%	+2.974	[+1.671, +4.632]	[+1.553, +4.921]
NVIDIA-Nemotron-Nano-9B-v2	71.71%	+3.303	[+1.868, +5.145]	[+1.711, +5.461]

The fixed offset-1 sensitivity estimates are +3.092 (97.5% interval [+1.579, +5.132]) for Llama and +3.487 ([+1.908, +5.658]) for Nemotron. Both familywise lower bounds exceed zero, yielding **ROBUSTNESS_SUPPORTED_BOTH_ARMS**. This extension was specified after the original Qwen14-to-Qwen7 outcome was known and is not an independent TEST replication.

Table S12 reports the deliberately bounded post-outcome extension. With the first-stage Qwen14 bank held fixed, the independent-pairing recovery opportunity is $113/3,800 = +2.974$ percentage points for Llama and $251/7,600 = +3.303$ points for Nemotron. Both familywise-corrected 97.5% query-cluster intervals lie entirely above zero, and both fixed offset-1 sensitivities remain positive. The frozen decision is **ROBUSTNESS_SUPPORTED_BOTH_ARMS**. Because the arms were specified after the primary outcome, share its TEST queries and Qwen14 stop bank, and use TP=1 alternative generation, they test only Gate-1 model-family persistence; they are neither independent confirmation nor a wider selector study.

16.3 Detailed related-work taxonomy

Table S13: Taxonomy of the closest allocation and verification families. The distinguishing object is the information available when the second action is selected, not whether a system happens to invoke more than one model.

Family	Decision point	Legal feedback	Main action	Gap addressed here
Query routing (Hu et al., 2024; Huang et al., 2025; Ong et al., 2025)	before response	prompt or preference features	choose one model	cannot condition on a realized weak-verifier stop
Response-conditioned cascades (Aggarwal et al., 2024; L. Chen et al., 2024; Dekoninck et al., 2025)	after a response	answer score or self-verification	accept or escalate	does not separate realized action availability from learnable action choice
Adaptive sampling and multi-model compute (Aggarwal et al., 2023; J. Chen et al., 2026; Ding et al., 2025)	before or during sampling	agreement, consistency, or predicted query difficulty	sample again or switch	does not audit finite-sample support before interpreting a learned selector
Code generation and verification (B. Chen et al., 2023, 2025; Liu et al., 2023; Ni et al., 2023)	after candidate execution	generated tests, public tests, or learned score	select, reject, or escalate a candidate	typically treats verification as selection evidence rather than measuring hidden recovery debt
This study	after a declared verifier stop for the prospective framework	outcome-blind stopped history only	resample or reroute, then stop	separates debt, support, and claim licensing; the all-episode audit is descriptive

16.4 Complete evidence-role table

This table retains the complete design and inferential-role entries summarized by the reading map in the main article. Its rows are separate applications, not a jointly completed three-gate experiment.

Table S14: Evidence map. Each dataset serves one declared gate role; sharing a response bank or TEST population does not create an independent replication. The rows are separate empirical applications of the framework; gates are not composed across datasets or rows.

Evidence	Gate role	Design	Inferential role	Licensed conclusion
Fresh MBPP+ Qwen pair	Gate 1	152 hash-split TEST queries; pinned generation and rescoring	prespecified primary	whether Base-only stopping debt is recoverable
MBPP+ matched-information audit	descriptive audit	nested query cross-fitting on all 1,520 episodes of the opened TEST bank	post-audit diagnostic	realized-maximum calibration, marginal fixed-action comparison, and no tested outcome-blind selector identified when to deviate from the fixed action
LiveCodeBench v5+v6	Gate 2 diagnosis	136 FIT, 68 DEV, one 137-query sealed TEST opening	separated frozen follow-up	support-shift diagnosis for the declared ladder
BigCodeBench v0.1.4	Gate 2 prospective	483 FIT, 211 unopened DEV, 405 unopened TEST; near-duplicate-group split	preregistered gate replication	frozen design misses two-sided count minima; no feature fit or outcome opening
Llama/Nemotron extension	robustness	same MBPP+ queries and Qwen14 stop bank; new alternative banks	post-outcome sensitivity	model-family persistence of the mechanism only
Historical 11-model archive	audit	fixed GSM8K, MATH-500, GPQA, and HumanEval+ replay	exploratory	why broad policy and static-routing claims are rejected

References

- Aggarwal, P., Madaan, A., Anand, A., Potharaju, S.P., Mishra, S., Zhou, P., ... Mausam (2024). AutoMix: Automatically mixing language models. *Advances in neural information processing systems* (Vol. 37, pp. 131000–131034). Retrieved from https://proceedings.neurips.cc/paper_files/paper/2024/hash/ecda225cb187b40ea8edc1f46b03ffda-Abstract-Conference.html
- Aggarwal, P., Madaan, A., Yang, Y., Mausam. (2023). Let’s sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 12375–12396). Retrieved from <https://aclanthology.org/2023.emnlp-main.761/>
- Chen, B., Zhang, F., Nguyen, A., Zan, D., Lin, Z., Lou, J.-G., Chen, W. (2023). CodeT: Code generation with generated tests. *Proceedings of the international conference on learning representations*. Retrieved from <https://openreview.net/forum?id=ktrw68Cmu9c> (arXiv:2207.10397)
- Chen, B., Zhu, M., Dolan-Gavitt, B., Shafique, M., Garg, S. (2025). Model cascading for code: A cascaded black-box multi-model framework for cost-efficient code completion with self-testing. *2025 international joint conference on neural networks (ijcnn)* (pp. 1–9). Retrieved from <https://doi.org/10.1109/IJCNN64981.2025.11227916>
- Chen, J., Xun, Z., Zhou, B., Qi, H., Zhang, H., Zhang, Q., ... Hu, S. (2026). Do we truly need so many samples? multi-LLM repeated sampling efficiently scales test-time compute. *Proceedings of the aaai conference on artificial intelligence* (Vol. 40, pp. 20083–20091). Retrieved from <https://ojs.aaai.org/index.php/AAAI/article/view/39094>
- Chen, L., Zaharia, M., Zou, J. (2024). FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, Retrieved from <https://openreview.net/forum?id=cSimKw5p6R>
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.d.O., Kaplan, J., ... others (2021). *Evaluating large language models trained on code*. arXiv. Retrieved from <https://arxiv.org/abs/2107.03374> (Technical report (not peer-reviewed))
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... Schulman, J. (2021). *Training verifiers to solve math word problems*. arXiv. Retrieved from <https://arxiv.org/abs/2110.14168> (Technical report (not peer-reviewed))
- Dekoninck, J., Baader, M., Vechev, M. (2025). A unified approach to routing and cascading for LLMs. *Proceedings of the 42nd international conference on machine learning* (Vol. 267, pp. 12987–13010). Retrieved from <https://proceedings.mlr.press/v267/dekoninck25a.html>
- Ding, D., Mallick, A., Zhang, S., Wang, C., Madrigal, D., Hipolito Garcia, M.D.C., ... Rühle, V. (2025). BEST-route: Adaptive LLM routing with test-time optimal compute. *Proceedings of the 42nd international conference on machine learning* (Vol. 267, pp. 13870–13884). Retrieved from <https://proceedings.mlr.press/v267/ding25d.html>
- Grattafiori, A., et al. (2024). *The Llama 3 herd of models*. arXiv. Retrieved from <https://arxiv.org/abs/2407.21783> (Technical report (not peer-reviewed))
- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., ... Steinhardt, J. (2021). Measuring mathematical problem solving with the MATH dataset. *Proceedings of the neural information processing systems track on datasets and benchmarks* (Vol. 1). Retrieved from <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html>
- Hu, Q.J., Bieker, J., Li, X., Jiang, N., Keigwin, B., Ranganath, G., ... Upadhyay, S.K. (2024). *RouterBench: A benchmark for multi-LLM routing system*. arXiv. Retrieved from <https://arxiv.org/abs/2403.12031> (ICML 2024 Agentic Markets Workshop

- (non-archival); arXiv:2403.12031)
- Huang, Z., Ling, G., Lin, Y., Chen, Y., Zhong, S., Wu, H., Lin, L. (2025). RouterEval: A comprehensive benchmark for routing LLMs to explore model-level scaling up in LLMs. *Findings of the association for computational linguistics: Emnlp 2025* (pp. 3860–3887). Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2025.findings-emnlp.208/>
- Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., ... Cobbe, K. (2024). Let’s verify step by step. *Proceedings of the international conference on learning representations* (Vol. 2024, pp. 39578–39601). Retrieved from https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html
- Liu, J., Xia, C.S., Wang, Y., Zhang, L. (2023). Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. *Advances in neural information processing systems* (Vol. 36, pp. 21558–21572). Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract.html
- Ni, A., Iyer, S., Radev, D., Stoyanov, V., Yih, W.-T., Wang, S., Lin, X.V. (2023). LEVER: Learning to verify language-to-code generation with execution. *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 26106–26128). Retrieved from <https://proceedings.mlr.press/v202/ni23b.html>
- NVIDIA (2025). *NVIDIA Nemotron Nano 2: An accurate and efficient hybrid Mamba-Transformer reasoning model*. arXiv. Retrieved from <https://arxiv.org/abs/2508.14444> (Technical report (not peer-reviewed))
- Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J.E., ... Stoica, I. (2025). RouteLLM: Learning to route LLMs from preference data. *Proceedings of the international conference on learning representations*. Retrieved from https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html
- Rein, D., Hou, B.L., Stickland, A.C., Petty, J., Pang, R.Y., Dirani, J., ... Bowman, S.R. (2024). GPQA: A graduate-level google-proof Q&A benchmark. *First conference on language modeling*. Retrieved from <https://openreview.net/forum?id=Ti67584b98>