When Local Choices Matter in Language Model Generation

Anonymous authors
Paper under double-blind review

Abstract

Language models can reach the same answer through different plans, and produce different answers from the same plan. To explain their errors and decide where to intervene, we need to know which choices influence the outcome. Differences between sampled answers mix the effect of a choice with randomness in later generation. We introduce outcome-resolution covariance (ORC), a matrix describing how future outcome probabilities vary across choices under a fixed continuation policy. Independent continuations separate choice effects from continuation noise and yield unbiased ORC estimates. We use one continuation to define intervention weights and another to predict their gain, separating effects of the shared choice from dependence within a continuation. Across three Qwen text-to-SQL panels, these forecasts more accurately predict gains on fresh continuations, including new questions and plans. On a 128-question panel, they reduce mean squared error by 80% relative to uncalibrated same-continuation forecasts; using them to allocate a fixed intervention budget offline improves correctness by 1.05 percentage points over equal-budget random selection. Recording outcome categories at different checkpoints reveals plan-specific failures that correctness scores hide and state differences that later computation removes. A covariance identity connects reweighting gains to probability shifts, branch variation, and target alignment. Ledger experiments show why alignment matters: on tasks with incorrect consensus targets, reweighting increases agreement while reducing correctness on average. Variation between branches provides the room for steering; alignment with the task determines the benefit.

1 Introduction

A paper airplane crosses a line on the floor. Was it the folding pattern, or a lucky throw? Each flight combines pattern effects with throw-to-throw variation. We compare patterns by independently making and throwing new planes, using the same paper type and launch protocol (Figure 1). Repeated trials separate differences between patterns from variation within each pattern. What we record also matters: two patterns may clear the line equally often and produce different landing distances. Recording distance distinguishes outcomes that a success label treats alike.

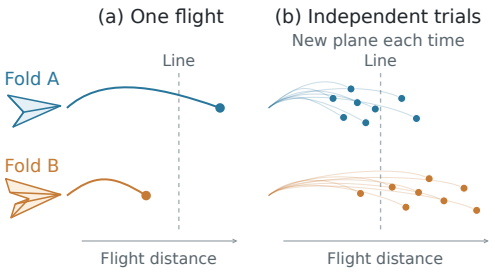

Figure 1: Independent trials separate folding-pattern effects from flight variability. Each trial uses a new plane; dots mark landing positions. Outcomes are illustrative.

This raises a general question: *when do local choices matter in language model generation?* A token or a planning block can send generation down different paths. Later steps may keep these paths apart or bring them back together, and the final answer may show the difference or hide it. Knowing which differences last tells us which decisions to examine when a model fails and where extra sampling or steering can change the result.

Each sampled answer combines choice effects with later randomness. The two cases in Figure 2 have the same marginals and token entropy, yet choices affect correct-

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

ness only in the second. Token entropy guides the selection of training tokens (Wang et al., 2025); rollouts are used to label reasoning steps by success (Wang et al., 2024) and assess the influence of tokens or sentences (Bigelow et al., 2025; Lin et al., 2025; Macar et al., 2026). Means computed from a few rollouts reflect both choice effects and continuation noise. Using the same continuation to define intervention weights and predict their gain also introduces dependence into the forecast. We freeze each choice and generate independent *replicas*. Their covariance equals the variance in success probabilities across choices. Using one replica to define weights and another to forecast gains isolates the effect that repeats on fresh continuations. Readouts at successive checkpoints show how this effect persists to the final answer (*temporal alignment*).

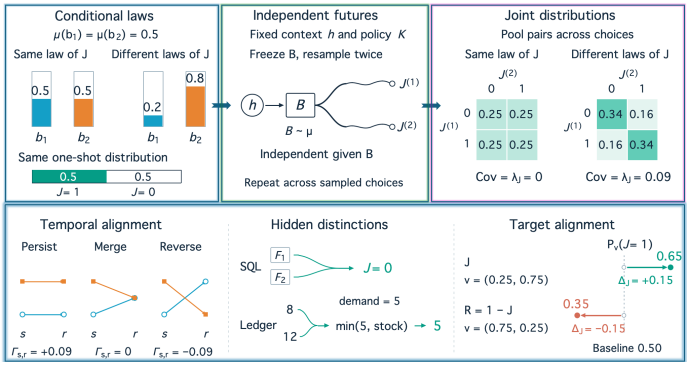

Figure 2: **Local choice effects revealed by independent continuations.** Bars show $\Pr(J = 1 \mid b)$; tables show joint probabilities of paired outcomes. Both cases have equal choice entropy and identical marginal correctness distributions. With context h and continuation policy K fixed, replicas sharing $B \sim \mu$ are independent given B and satisfy $\mathrm{Cov}(J^{(1)}, J^{(2)}) = \mathrm{Var}_{B\sim\mu}[\Pr(J = 1 \mid B)] = \lambda_J$, where λ_J is the correctness capacity. Bottom: temporal alignment of conditional means, distinctions merged by scoring or computation, and target alignment. SQL categories F_1 and F_2 denote original-database and suite-only mismatches, respectively. For the $0.2/0.8$ case, reweighting toward J or $R = 1 - J$ changes correctness from 0.50 to 0.65 or 0.35, respectively. Weight pairs are ordered as (b_1, b_2). Values are illustrative.

With several outcome categories, this covariance becomes a matrix, which we call *outcome-resolution covariance* (ORC; Equation (1)). ORC compares alternatives at a fixed prefix under a fixed continuation policy. For a scalar score, its *capacity* measures variation across choices. A gain forecast estimates the effect of a particular reweighting; target alignment determines how that effect translates into task benefit (Equation (5)). We use independent replicas in SQL and ledgers to measure these relationships and reveal distinctions hidden by correctness. For example, a Spider plan hard-codes Europe as continent identifier 2: its query passes on the original database and fails when a test database changes the identifier (Appendix L).

Contributions.

1. **Independent continuations identify choice effects.** We use pick–freeze sampling (Sobol', 1990; Gamboa et al., 2014; Janon et al., 2014) to estimate ORC without bias. Uncorrected SQL capacity estimated with two continuations per plan is 1.57 times the independent corrected reference on Qwen and 1.42 on Gemma; correction also changes question rankings.

2. **Independent forecasts connect measurement to intervention.** Once weights are fixed, an independent continuation gives a conditionally unbiased forecast of their gain. Across three Qwen SQL panels, mean signed errors range from -0.24 to 0.30 pp, versus 2.0–2.1 pp overprediction with continuation reuse. We use these forecasts to allocate a fixed intervention budget.

3. **Outcome resolution distinguishes scoring from computation.** Categorical SQL outcomes reveal repeatable plan-specific failures even when all sampled queries fail. In ledgers, the shipment calculation gives the correct action for 70 continuations with an incorrect inventory state.

4. **Target alignment connects capacity to task benefit.** Ledger capacity correlates more closely with held-out target-agreement gains ($r = 0.80$) than correctness gains ($r = 0.30$). On the five tasks with an incorrect consensus target, reweighting increases agreement and reduces correctness on average (Equation (5)).

2 OUTCOME-RESOLUTION COVARIANCE

Fix a prompt x, an autoregressive policy, and a finite horizon T with absorbing EOS. At prefix $h = H_t = (x, Y_{1:t})$, let $B \sim \mu$ be a token or a block ending at a bounded stopping time, such as a completed plan. The continuation policy $K(\cdot \mid h, b)$ generates the remaining text. Natural sampling or a specified reference distribution determines the choice law μ.

Let $Z = \Phi(H_T) \in \{1, \dots, d\}$ be an observed outcome, including any failure categories, encoded as $X = \boldsymbol{e}_Z$. The conditional future law v_b, its average q, and *outcome-resolution covariance* (ORC) are

$$v_b = \mathbb{E}_K[X \mid h, b], \qquad q = \mathbb{E}_\mu v_B, \qquad \Sigma = \mathrm{Cov}_\mu(v_B). \tag{1}$$

All quantities below are conditional on h. Each choice starts a *branch* of possible continuations. The matrix describes how outcome probabilities differ across branches: its diagonal measures variation in each probability, and its off-diagonal entries show which probabilities rise or fall together. A scalar score $R = c^\top X$ has capacity $\lambda_R = c^\top \Sigma c = \mathrm{Var}_\mu \mathbb{E}_K[R \mid h, B]$; total categorical capacity is $\mathrm{tr}\,\Sigma$. ORC is zero exactly when every supported choice has the same categorical future law (Proposition A.1).

Which distinctions does a readout keep? Any linear readout AX has ORC

$$\Sigma_{AX} = A\Sigma A^\top. \tag{2}$$

This includes merging several outcome categories into one score. Two plans can produce different execution failures at the same correctness probability. The categorical matrix preserves this distinction, which the correctness readout merges. One-hot encoding a tuple, such as a ledger state and action, preserves its joint distribution; a vector of feature scores represents their conditional means. Appendix G specifies when a representation determines the conditional outcome law.

How much uncertainty does the choice resolve? Total covariance separates variation across choices from variation within their continuation distributions:

$$\underbrace{\mathrm{Diag}(q) - qq^\top}_{C = \mathrm{Cov}(X)} = \Sigma + \mathbb{E}_\mu \mathrm{Cov}_K(X \mid B), \qquad \rho_R = \frac{\lambda_R}{\mathrm{Var}_{\mu K}(R)} \in [0, 1]. \tag{3}$$

For a positive denominator, ρ_R is the fraction of outcome variance resolved by the choice. This is the conditional-variance decomposition underlying variance-component and return-variance analyses (Sobel, 1982; Tamar et al., 2013; Sobol', 1990). It gives an interpretable scale for comparing capacity across outcomes and models. Appendix B connects ρ_R for binary outcomes and normalized categorical ORC to dependence measures.

3 TEMPORAL AND TARGET ALIGNMENT

Which distinctions persist across checkpoints? For checkpoint readouts X_s, write $m_s(b) = \mathbb{E}_K[X_s \mid h, b]$ and stack these means into $V(B)$. The joint checkpoint response matrix is

$$\Gamma = \mathrm{Cov}_\mu(V(B)), \qquad \Gamma_{s,r} = \mathrm{Cov}_\mu(m_s(B), m_r(B)). \tag{4}$$

We use the same checkpoint definitions across continuations: a bounded token count or an event such as a completed draft. Diagonal blocks give ORC at each checkpoint; off-diagonal blocks measure *temporal alignment* across checkpoints. For scalar events, a positive entry means choices favoring the earlier event also tend to favor the later one. Equal marginal capacities can accompany very different dynamics. For independent fair bits U, V revealed by a block, (U, U), (U, V), and $(U, 1-U)$ share diagonal entries $1/4$, with respective cross-entries $1/4$, 0, and $-1/4$. With two choices and nonzero earlier capacity (Figure 2, bottom left), a zero cross-entry means the later conditional means are equal.

When does a responsive choice improve the target? Reweight the alternatives from μ to $\nu \ll \mu$ while retaining K. Put $m_R(b) = \mathbb{E}_K[R \mid h, b]$, $w = \nu/\mu$, and $\Delta_R = \mathbb{E}_{\nu K} R - \mathbb{E}_{\mu K} R$. The covariance identity factors the response into probability shift, capacity, and alignment; Cauchy–Schwarz gives its bound:

$$\Delta_R = \sqrt{\chi^2(\nu\|\mu)\lambda_R}\,\alpha_R(w-1), \qquad |\Delta_R| \le \sqrt{\chi^2(\nu\|\mu)\lambda_R}, \tag{5}$$

where $\alpha_R(w-1)$ is the correlation of $w-1$ with $m_R(B)$, defined as zero if either variance vanishes. Thus perturbation size, outcome-specific capacity, and target alignment jointly determine the response. In Figure 2, $\chi^2(\nu\|\mu) = 0.25$, $\lambda_J = 0.09$, and $\alpha_J = \pm 1$ give $\Delta_J = \pm 0.15$. A direction learned for a proxy target R has its own alignment with task correctness J. Measuring both on independent continuations shows how a gain in R carries over to J. Appendices C.3, D and F give the local Fisher form, sequential responses, and a policy-value decomposition when feedback changes K.

4 Estimating ORC with Independent Replicas

Generate independent continuations of the same choice. Draw $m \geq 2$ iid blocks $B_i \sim \mu$ and $n \geq 2$ independent suffixes per block, with output vectors $X_{i1}, \ldots, X_{in}$. Two suffixes of one block are independent given the block, so $\mathrm{Cov}(X_{i1}, X_{i2}) = \Sigma$, the matrix form of $\mathrm{Cov}(J^{(1)}, J^{(2)}) = \lambda_J$ in Figure 2. Let $\bar{X}_i$ and S_i be the within-block sample mean and unbiased sample covariance, and $\bar{X} = m^{-1}\sum_i \bar{X}_i$. Using the pick–freeze design (Sobol', 1990; Janon et al., 2014), we estimate ORC by

$$\widehat{\Sigma}_U = \underbrace{\frac{1}{m-1}\sum_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^\top}_{\text{observed spread of block means}} - \underbrace{\frac{1}{mn}\sum_i S_i}_{\text{suffix measurement noise}} . \tag{6}$$

It is unbiased for Σ. For a scalar readout, the uncorrected first term has expectation $\lambda_R + \mathbb{E}_B \mathrm{Var}_K(R \mid B)/n$, so its noise contribution depends on the continuation budget. On a fixed library of m equally weighted choices, multiplying (6) by $(m-1)/m$ estimates that library's capacity. The original two-replica panels use covariance across replicas with denominator $m-1$. We retain both positive and negative estimates (Appendix E.1).

Applying this design to stacked checkpoint vectors identifies Γ. Readouts at checkpoints s and r of the same continuation have covariance

$$\mathrm{Cov}(X_s, X_r) = \Gamma_{s,r} + \mathbb{E}_B \mathrm{Cov}_K(X_s, X_r \mid B). \tag{7}$$

The first term comes from the shared choice; the second captures dependence within a continuation.

Forecast protocol. Use one continuation per choice to define weights and an independent one to forecast their gain. Conditional on the choices and the data used to set the weights, it is unbiased for the expected gain on fresh continuations (Equations (66) and (67)). The same two continuations provide both forecasts. Section 5.1 tests them on fresh outcomes using paired task or database resampling; Appendix E gives estimator variance and sampling-cost tradeoffs.

5 Local Choices in Executable Computation

We use SQL query planning to test independent gain forecasts and budget allocation, and ledger evidence selection to test capacity-guided selection and target alignment. A ledger block selects the effective event revisions; a reference calculator checks the resulting inventory and shipment action (Appendix H.4). Qwen3.8-27B is the primary model; we repeat the original SQL experiment with Gemma-4-31B-it. Both sample from the full softmax at temperature one with thinking disabled. Generation resumes from saved block token IDs with independent suffix seeds. We center estimates within tasks and weight tasks equally. For 95% intervals, we resample ledger tasks within strata or whole SQL databases (Appendix H).

5.1 Independent forecasts of reweighting gains

The original SQL panel contains 32 hard or extra-hard Spider questions across 19 databases, selected by query structure before generation (Yu et al., 2018). For each question and model, we sample 16 plans and two independent continuations per plan. Each continuation produces a draft query and a final query after review; execution against the distilled test suite determines both outcomes (Zhong et al., 2020). Qwen's draft–final covariance contains similarly sized shared-plan and continuation components (Figure 3, top).

Table 1: **Independent forecasts generalize to new questions and plans.** Each panel has 32 questions and 16 plans per question. Forecasts s, c and independent gain e are in pp; L is in pp^2, with 95% database-cluster intervals and equal question weighting.

Model	Panel	Same s	Cross c	Gain e	Loss difference L
Qwen	Original	4.485	2.204	2.439	8.574 [3.149, 15.527]
Qwen	**New questions**	5.180	3.116	3.106	10.367 [3.813, 17.709]
Gemma	Original	0.911	0.719	0.806	0.128 [−0.197, 0.530]

Freeze a reweighting, then test its prediction. For plan i and replica $r \in \{1, 2\}$, let D_{ir}^{freeze} be the recorded draft test-suite label and F_{ir} the final correctness label verified by execution (Appendix I). We fix the weights at $\pi_{ir} = 2^{D_{ir}^{\text{freeze}}} / \sum_j 2^{D_{jr}^{\text{freeze}}}$. With $a_{ir} = \pi_{ir} - 1/16$, the same-continuation forecast is $s_r = \sum_i a_{ir} F_{ir}$; the independent forecast uses the other continuation, $c_r = \sum_i a_{ir} F_{i,3-r}$. We evaluate these fixed directions using fresh continuations of the same plans: $e_r = \sum_i a_{ir} \bar{F}_{i,E}$. We compare the two forecasts using

$$L = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{2} \sum_{r=1}^{2} \left[(s_{qr} - e_{qr})^2 - (c_{qr} - e_{qr})^2 \right]. \quad (8)$$

Positive L means lower squared prediction error for the independent forecast. The comparison includes all questions and constant directions. Gains are in percentage points (pp) and L in pp^2.

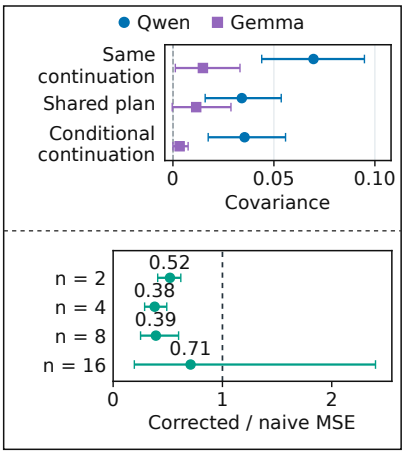

Figure 3: **Independent continuations reveal the contribution of a shared choice.** Top: draft–final covariance and its plan and continuation components on 32 SQL questions across 19 databases. Bottom: corrected/naive MSE against an independent reference on 60 math prefixes, with 20 candidate tokens per prefix; n is the number of suffixes per candidate and the dashed line marks equal error. Bars show 95% database- and problem-cluster intervals, respectively.

Forecasts also improve on new questions and plans. Eight fresh continuations per original plan yield a Qwen gain of 2.44 pp, compared with same-continuation and independent forecasts of 4.48 and 2.20 pp (Table 1). On 32 additional questions, we sample 16 new plans per question, with two discovery and four evaluation continuations each. The gain is 3.11 pp, and the forecasts are 5.18 and 3.12 pp: independent prediction reduces MSE by 77%. The interval for L is positive on these two panels (Figure 4 and Appendix I.7). For Gemma, draft labels yield uniform weights for 55 of 64 directions, including both on 27 of 32 questions. These directions have $\chi^2(\nu \| \mu) = 0$ and zero predicted and measured gains (Equation (5)). Gemma's plans have positive final capacity (Table 3); its loss contrast is 0.128 [−0.197, 0.530] pp^2. Qwen's result also holds on subsets defined by input structure (Appendix I.2).

Continuation reuse adds draft–final dependence. Given the draft labels used to set the weights, the other replica and fresh continuations have the same conditional expected gain. The same-continuation forecast also includes draft–final dependence within that continuation (Appendix I.3). On fresh continuations of Qwen's original plans, this component accounts for 43.8% of pooled covariance; it is smaller for Gemma.

Noise correction makes the capacity scale comparable across budgets. With two continuations per plan, uncorrected SQL capacity is 1.566 times the independent corrected reference on Qwen and 1.417 on Gemma. Correction changes question rankings (Table 13) and keeps the mean over continuation subsets constant across budgets (Appendix I.5). On 30 math problems (60 prefixes, top-20 token candidates), correction with four continuations per candidate reduces capacity-estimation MSE against an independent reference by 62% (Figure 3, bottom; Appendix H.2).

Table 2: **Independent forecasts and budgeted SQL allocation.** We average forecast errors over all 128 potential interventions. Policy gains compare interventions on 32 selected questions with uniform plan sampling, using the same denominator of 128 questions. Brackets give 95% database-cluster intervals.

Quantity	SAME	CROSS
Forecast MSE (pp^2)	15.61 [9.66, 23.63]	3.15 [1.72, 5.04]
Forecast bias (pp)	2.08 [1.56, 2.72]	0.30 [−0.07, 0.73]
Policy gain (pp)	2.11 [1.22, 2.96]	2.33 [1.57, 3.05]

Using forecasts to allocate a fixed intervention budget. The third Qwen panel has 128 new questions from 13 previously used databases: 16 plans each, with two discovery and eight evaluation continuations. Oracle draft test-suite feedback defines KL-0.05 reweightings; selectors choose 32 of 80 questions with nonconstant weights. CROSS ranks independently predicted gains; ORC-CAPACITY ranks corrected variance. We evaluate both on fresh continuations of the fixed plans. Across all 128 potential interventions, SAME forecasts 5.30 pp and CROSS 3.51 pp, against a fresh gain of 3.21 pp. CROSS reduces MSE by 79.8%; calibration on earlier panels gives MSEs of 6.67, 3.12, and 3.45 pp^2 for SAME, CROSS, and POOLED (Appendix M.3). CROSS allocation gains 2.33 pp over uniform plan sampling; random allocation gains 1.29 pp. With directions fixed, ranking independent forecasts adds 1.05 [0.52, 1.49] pp over random. CROSS's estimated gain differences from SAME, NAIVE, and ORC-CAPACITY are 0.22 [−0.35, 0.78], 0.10 [−0.16, 0.31], and 0.06 [−0.15, 0.22] pp, respectively (Tables 2 and 27). CROSS and SAME share 25 of their 32 selections. The original two-direction test gives $L = 9.60$ [5.31, 15.97] pp^2; Appendix M gives full comparisons and reference-SQL overlap checks.

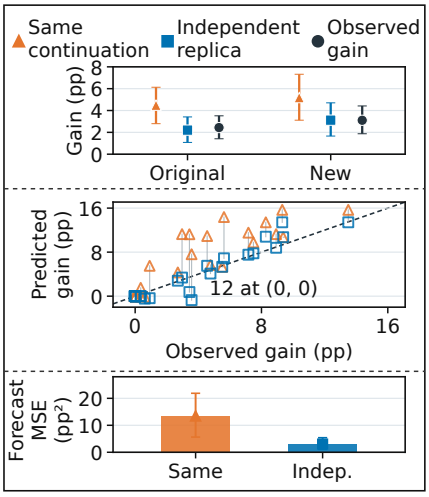

Figure 4: **Independent forecasts track gains on new questions and new plans.** Top: forecasts and fresh-continuation gains on two Qwen panels. Middle: means over both directions for each new question; 12 fall at zero. Bottom: direction-level prediction MSE, 13.41 versus 3.04 pp^2 (77% reduction). Bars show 95% database-cluster intervals.

5.2 Plan effects across checkpoints

Both models have positive final SQL capacity (Figure 5 and Table 3). Qwen's positive draft–final covariance shows that plans favoring correct drafts also favor correct finals in separate continuations. Its ratio of panel-mean capacity to variance, $\rho_F = \lambda_F / v_F = 0.461$ [0.264, 0.647], attributes 46% of measured *within-question* variance to plans under the stated policy. Both capacity and covariance remain positive on Qwen's input-defined subset (Appendix K). In the ledger follow-up, blocks favoring correct early inventory also favor correct final actions: $\Gamma_{EL} = 340$ [133, 554] pp^2.

5.3 Outcome resolution reveals branch distinctions

Categorical outcomes reveal plan-specific failure modes. The categorical SQL readout distinguishes execution failure, disagreement on the original database, disagreement appearing only on the expanded suite, and success. In `world_1_0766`, all 32 sampled finals fail the expanded suite: three plans fail on the original database in both continuations, and thirteen fail only on the expanded tests in both (Figure 5, bottom). On these two coordinates,

$$\widehat{\Sigma}_F = \begin{pmatrix} 1625 & -1625 \\ -1625 & 1625 \end{pmatrix} \text{ pp}^2, \qquad \widehat{\lambda}_{\text{correct}} = 0. \tag{9}$$

Table 3: **Natural branch capacity and temporal alignment.** Each task has 16 blocks and two suffixes per block. Entries use the original cross-column estimator, in pp^2 (10^4 times probability-squared units); v is the matching within-task variance budget. We obtain intervals by resampling ledger tasks or SQL databases. We use Qwen for the follow-up with early records.

		Qwen		Gemma	
Task panel	Quantity	Estimate	95% interval	Estimate	95% interval
Ledger: 24 tasks	Joint correctness λ	361	$[108, 622]$	21	$[-12, 76]$
	Variance budget v	1507	—	177	—
	Early–late Γ_{EL}	340	$[133, 554]$	—	—
SQL: 32 questions, 19 databases	Final capacity λ_F	384	$[188, 594]$	184	$[11, 394]$
	Draft–final Γ_{DF}	341	$[160, 536]$	115	$[-3, 287]$
	Variance budget v_F	833	—	340	—

Independent pairs reveal reproducible plan-dependent failure categories. The complementary failure probabilities give a negative off-diagonal entry whose magnitude measures variation hidden by correctness scores.

Across the full Qwen panel, "original pass, expanded-suite failure" has final capacity 352 $[160, 602]$ pp^2 and draft–final covariance 335 $[140, 601]$ pp^2, and its capacity remains positive after excluding the two `world_1` questions. Among final queries that pass on the original database, 239/843 on Qwen and 172/922 on Gemma fail expanded tests.

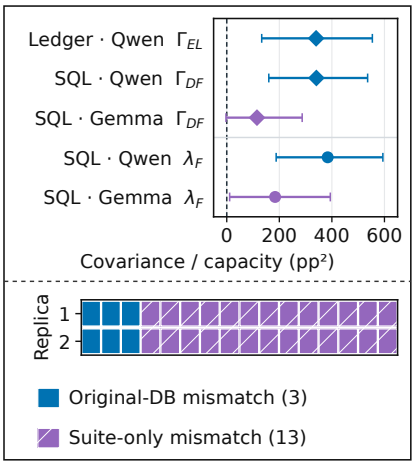

Figure 5: **Temporal and categorical readouts reveal branch structure.** Top: ledger and SQL checkpoint covariance and SQL final capacity, with 95% task- or database-cluster intervals, in pp^2. Bottom: all sampled queries fail for Qwen on `world_1_0766`. Two independent continuations from each of 16 plans reproduce the same failure category: 3 plans fail on the original database and 13 fail only on the expanded suite. Columns are plans grouped by category; rows are continuations.

Different states can also yield the same correct action. In the Qwen ledger follow-up, we record the early inventory balance, final state, and shipment. Among valid completions with interpretable early records, replay reproduces 743/743 final states and 742/743 actions. In 70 continuations an incorrect state yields the correct shipment, because under $\min(\text{demand}, \text{allocable})$ different stocks above demand produce the same shipment. A distinction can therefore survive in the reported state and disappear during computation.

5.4 From branch capacity to task-aligned response

On 48 new Qwen ledger tasks, we use consensus among eight complete answers to set a fixed target state and action from the model's own outputs. Each task has eight natural blocks, four discovery continuations per block to learn an exponential tilt and rank tasks by target-agreement capacity, and eight fresh continuations per block to evaluate frozen policies. Four strata combine two treatments of a voided event—removal or fallback to its latest posted revision—with whether omitting a designated event changes the action or only the state. Each selector chooses three tasks per stratum at KL 0.05 for nonconstant directions and zero otherwise. Joint correctness J is the primary outcome; target agreement R and comparisons at equal total KL are secondary (Appendix H.6).

Capacity transfers across samples. Capacity estimates from the discovery and evaluation samples correlate at $r = 0.818$ $[0.705, 0.892]$. Discovery capacity correlates more closely with gains in target agreement ($r = 0.796$ $[0.690, 0.872]$) than correctness gain ($r = 0.298$ $[-0.120, 0.677]$). Of 42

Table 4: **Ledger response: forecasting a fixed intervention and choosing where to intervene.** Gains are in percentage points, with 95% stratified-task intervals. R measures agreement with the fixed target; J measures joint correctness of the state and action.

A. Fixed task directions, requested KL 0.05 per task

Outcome	Discovery forecast	Independent evaluation
Target agreement R	5.842 [4.883, 6.791]	3.248 [2.371, 4.161]
Joint correctness J	4.064 [2.762, 5.289]	2.047 [0.995, 3.079]

B. Allocation comparisons: ORC minus comparator

Comparator	Total KL	Target agreement R	Joint correctness J
Random selection	0.525	1.023 [0.287, 1.835]	0.270 [−0.471, 1.047]
Random, nonconstant	0.600	0.908 [0.190, 1.694]	0.208 [−0.504, 0.955]
Block surprisal	0.500	0.982 [0.249, 1.847]	0.287 [−0.354, 1.025]
Outcome uncertainty	0.600	0.276 [−0.475, 1.079]	0.276 [−0.475, 1.079]
Naive / predicted response	0.600	0.111 [−0.050, 0.382]	0.111 [−0.050, 0.382]
Diffuse, matched KL	0.600	0.039 [−0.560, 0.669]	−0.359 [−0.922, 0.218]

Panel A averages all 48 tasks, including six constant directions. In panel B, selection policies choose three tasks per stratum; ORC uses total KL 0.600. We average random selection over its selection law. Diffuse allocation spreads KL 0.600 over all 42 nonconstant directions. Two rows have equal R and J contrasts because their selections differ from ORC's only on correct-target tasks, where $R = J$.

nonconstant directions, 37 improve target agreement in the evaluation pool. Across all 48 tasks, discovery and evaluation gains average 5.842 and 3.248 pp, respectively (Table 4A).

Target quality shapes task utility. ORC's target-agreement gain over random exceeds its joint-correctness gain by 0.753 $[0.101, 1.583]$ pp (Table 4B). Correct-target tasks contribute 0.583 pp to both gains. The five wrong-target tasks contribute $+0.441$ $[0.063, 0.917]$ pp to agreement and -0.313 $[-0.713, -0.010]$ pp to correctness: stronger agreement with a mistaken target lowers task success. The gap persists across all five leave-one-task-out calculations (Appendix H.10).

Allocation at a common KL budget. ORC also improves agreement over random selection at equal total KL. Selecting by naive variance or predicted response chooses 11 of ORC's 12 tasks and gives similar gains (Table 4B). In a post hoc comparison, spreading KL 0.600 over all 42 nonconstant directions improves joint correctness over uniform blocks by 1.141 $[0.564, 1.717]$ pp; the ORC-minus-diffuse contrast is -0.359 $[-0.922, 0.218]$ pp.

Capacity and uncertainty predict response. In a post hoc analysis, we use four-fold cross-fitting to predict R and J gains at KL 0.05 and corrected R capacity. All models use mean agreement q. Adding corrected capacity improves R-gain prediction. Adding naive variance instead gives lower MSE estimates, by 2.231 $[0.174, 6.921]$ pp^2 for J. Corrected capacity measures branch variation, while naive variance also includes variation within branches. The model using q and uncertainty $q(1-q)$ has the lowest MSE estimate for J gain (Appendix I.6).

Four of the five wrong targets have the correct shipment and an incorrect inventory state. This pattern is consistent with the many-to-one computation that maps different state errors to the same action during evaluation. The responses of R and J to a reweighting depend on $\mathrm{Cov}(w-1, m_R)$ and $\mathrm{Cov}(w-1, m_J)$. In the incorrect-target group, opposite covariance signs explain the observed average increase in agreement and decrease in correctness.

Source feedback changes the continuation policy. A checker identifies discrepancies between records and source events in a new turn, changing K. After cohort standardization, the estimated gain in Qwen's joint correctness is 5.595 pp over ordinary continuation (95% interval $[0.671, 10.518]$ pp; nested repair-prefix resampling). Appendices H.7 and J specify the deterministic repair cohort, standardization, and replay results.

6 Related Work

Statistical foundations. Sensitivity analysis studies conditional-mean variances (Sobol', 1990) and covariances (Gamboa et al., 2014), estimated by conditionally independent pick–freeze replicas (Janon et al., 2014). Return-variance recursions track trajectory variation (Sobel, 1982; Tamar et al., 2013). We use covariance across choices to connect outcome categories and checkpoints to predictions of intervention gains.

Resampling reasoning paths. Forking Paths studies how outcome distributions change along trajectories and with alternative tokens (Bigelow et al., 2025). Thought Anchors and Thought Branches measure causal influence through sentence resampling (Bogdan et al., 2025; Macar et al., 2026). Critical Tokens Matter uses rollouts and contrastive estimation to identify and weight failure-related tokens in preference training (Lin et al., 2025). ORC measures how conditional outcome laws vary across choices: $\mathbb{E}\|v_B - q\|^2 = \operatorname{tr}\Sigma$ for a fixed prefix and continuation policy. Independent replicas correct for noise from finite continuations and predict how fixed reweightings will affect fresh continuations.

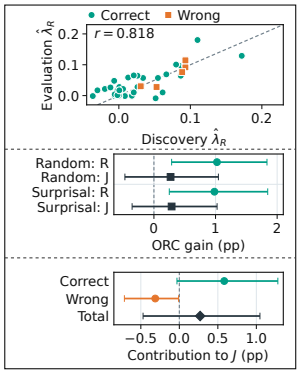

Figure 6: **Capacity and target alignment in ledger tasks.** Top: discovery versus independent evaluation capacity on 48 tasks; circles mark correct targets, and squares incorrect targets. Middle: ORC's gains over random and block-surprisal selection in target agreement R and joint correctness J. Bottom: contributions of correct and incorrect targets to the full-panel correctness gain over random selection: $+0.583 - 0.313 = +0.270$ pp. Intervals use paired, stratified task resampling.

Process values and target response. Math-Shepherd labels steps by rollout success frequency or whether any rollout succeeds (Wang et al., 2024). For the same choices, policy, and reward, λ_R measures variance across the underlying success probabilities. ORC corrects the estimation noise from finite rollouts. Process supervision also uses human step judgments (Lightman et al., 2024). Process advantage verifiers give a small-step policy-improvement lower bound with additive prover-advantage variance and prover/base alignment terms (Setlur et al., 2025). Equation (5) gives the exact response to finite reweighting under a fixed continuation policy. Independent continuations measure capacity and target alignment. The local form connects to score identities and Fisher geometry (Williams, 1992; Kakade, 2001). Entropy-based training computes its loss on high-entropy token positions (Wang et al., 2025); adaptive bandit methods allocate compute across queries (Zuo & Zhu, 2026). We allocate fixed interventions offline using independent gain forecasts.

7 When a Local Choice Matters

A local choice matters when it changes outcome probabilities under a fixed continuation policy. ORC measures these changes, and checkpoint readouts track them through scoring and computation. Task benefit depends on alignment with success: steering toward wrong ledger targets raises agreement and lowers correctness on average. Across three Qwen SQL panels, mean signed forecast errors range from -0.24 to 0.30 pp with independent continuations, versus 2.0–2.1 pp with continuation reuse. Allocating a fixed budget with these forecasts improves correctness over random selection on the 128-question panel.

Limitations and Future Work

We used Qwen for the SQL follow-ups and ledger allocation, and Gemma for the original SQL panel. We test finite sets of choices on tasks with executable checks, using ground-truth SQL feedback and ledger targets from model consensus. Larger samples would clarify whether CROSS outperforms SAME in allocation, independent continuations improve Gemma forecasts, and ORC outperforms random selection in ledger correctness. Future work will study online interventions without labels, better target construction, and adaptive sampling across more tasks, models, and databases.

AI USE STATEMENT

AI tools supported the work, including polishing the writing, analyzing experimental results, and checking and correcting mathematical formulations. The authors led the problem formulation and description, experimental design and validation, theoretical analysis, and preparation of the initial and final manuscripts, and take responsibility for the complete work.

ETHICS STATEMENT

This work studies language-model generation using public mathematical and text-to-SQL benchmarks and synthetic inventory ledgers. The experiments involve model-generated outputs and executable task evaluations, with no recruitment of human participants. Our findings on target alignment show why intervention objectives should be checked and human oversight retained when steering methods are applied beyond these controlled tasks.

REPRODUCIBILITY STATEMENT

The appendix defines the outcomes, sampling designs, estimators, resampling units, and generation and scoring protocols. The supplementary package contains anonymous labels for each discovery and evaluation continuation in all three SQL panels, the fixed intervention weights and selections, and analysis code that runs on a CPU. The code uses these labels to recompute the primary forecasts, held-out gains, and database-cluster confidence intervals, repeat the continuation-budget analysis, and regenerate the corresponding figures and tables.

REFERENCES

James N. Arvesen. Jackknifing U-statistics. *The Annals of Mathematical Statistics*, 40(6):2076–2100, 1969. doi: 10.1214/aoms/1177697287.

Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking paths in neural text generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2b6a8ca3d06ffa2e044bff8c723863ff-Abstract-Conference.html.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which LLM reasoning steps matter? *arXiv preprint arXiv:2506.19143*, 2025. URL https://arxiv.org/abs/2506.19143.

Andreas Buja. Remarks on functional canonical variates, alternating least squares methods and ACE. *The Annals of Statistics*, 18(3):1032–1069, 1990. doi: 10.1214/aos/1176347739.

Fabrice Gamboa, Alexandre Janon, Thierry Klein, and Agnès Lagnoux. Sensitivity analysis for multidimensional and functional outputs. *Electronic Journal of Statistics*, 8(1):575–603, 2014. doi: 10.1214/14-EJS895. URL https://doi.org/10.1214/14-EJS895.

Alexandre Janon, Thierry Klein, Agnès Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two Sobol index estimators. *ESAIM: Probability and Statistics*, 18: 342–364, 2014. doi: 10.1051/ps/2013040.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems 14*, pp. 1531–1538. MIT Press, 2001.

H. O. Lancaster. The structure of bivariate distributions. *The Annals of Mathematical Statistics*, 29 (3):719–736, 1958. doi: 10.1214/aoms/1177706532.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhances LLM's reasoning capability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37906–37918. PMLR, 2025. URL https://proceedings.mlr.press/v267/lin25j.html.

Uzay Macar, Paul C. Bogdan, Senthooran Rajamanoharan, and Neel Nanda. Thought branches: Interpreting LLM reasoning requires resampling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/954d303839bf7fc458acea7384322255-Abstract-Conference.html.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–606, 1934. doi: 10.2307/2342192.

Giorgos Nikolaou, Tommaso Mencattini, Donato Crisostomi, Andrea Santilli, Yannis Panagakis, and Emanuele Rodolà. Language models are injective and hence invertible. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5f124a1979570872ff46a6de8fbac35c-Abstract-Conference.html.

Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10(3–4):441–451, 1959. doi: 10.1007/BF02024507.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/98711dea460bdefe0e651ca23ec98ba2-Abstract-Conference.html.

Matthew J. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982. doi: 10.2307/3213832.

I. M. Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990. URL https://www.mathnet.ru/eng/mm2320. In Russian.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Temporal difference methods for the variance of the reward to go. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28(3) of *Proceedings of Machine Learning Research*, pp. 495–503. PMLR, 2013. URL https://proceedings.mlr.press/v28/tamar13.html.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Learning the variance of the reward-to-go. *Journal of Machine Learning Research*, 17(13):1–36, 2016. URL https://jmlr.org/papers/v17/14-335.html.

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

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3850. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a797c2d2e0c1fdabf4d1ab8cd0b465c6-Abstract-Conference.html.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3911–3921. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1425. URL https://aclanthology.org/D18-1425/.

Ruiqi Zhong, Tao Yu, and Dan Klein. Semantic evaluation for text-to-SQL with distilled test suites. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 396–411. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.29. URL https://aclanthology.org/2020.emnlp-main.29/.

Bowen Zuo and Yinglun Zhu. Strategic scaling of test-time compute: A bandit learning approach. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d8d2c5f79c53a2a685dbdf93f78c5695-Abstract-Conference.html.

A Distributional Geometry and Observable Resolution

Throughout this appendix, we assume finite horizons, a finite vocabulary with absorbing EOS, and bounded observation vectors. Each estimand is defined by a fixed prefix, block law, observation map, and continuation policy. Expectations use the joint law of blocks and continuations, whether blocks follow the natural or a chosen reference law. Sequential budgets follow a single autoregressive process throughout the window.

Write $H_t = (x, Y_{1:t})$ and $\mathcal{F}_t = \sigma(H_t)$. Under policy p, a natural k-token block at h has law

$$\mu(b \mid h) = \prod_{j=1}^{k} p(b_j \mid h, b_{<j}). \tag{10}$$

A reference law assigns weights to specified alternatives. Under deterministic decoding, the natural law concentrates on one choice. Each study specifies its choice law, continuation policy, and observation map.

A.1 Complete outcome laws, projections, and coarsening

Proposition A.1 (Branch information and observable projections)**.** *For independent $B, B' \sim \mu$,*

$$\Sigma = \tfrac{1}{2}\mathbb{E}[(v_B - v_{B'})(v_B - v_{B'})^\top]. \tag{11}$$

Thus $\Sigma = 0$ exactly when $Z \perp B$ under the joint law μK: every supported choice has the same categorical future law. For any fixed linear readout AX, its ORC is $A\Sigma A^\top$, as in (2)*. For a categorical coarsening f, the readout matrix has entries $A_{u,z} = \mathbf{1}\{f(z) = u\}$.*

Proof of Proposition A.1. For iid $v_B, v_{B'}$ with mean q, expansion gives

$$\mathbb{E}[(v_B - v_{B'})(v_B - v_{B'})^\top] = 2\{\mathbb{E}[v_B v_B^\top] - qq^\top\} = 2\Sigma.$$

The diagonal entries of a zero covariance matrix show that each coordinate of v_B is almost surely constant. Since v_b is the entire conditional categorical law, this is equivalent to $\Pr(Z = z \mid B) = \Pr(Z = z)$ for each z, or $Z \perp B$. For a fixed matrix A, linearity of conditional expectation gives $\mathbb{E}[AX \mid B] = Av_B$, and taking covariance proves (2). □

If f maps fine labels to coarse labels, define $A_{u,z} = \mathbf{1}\{f(z) = u\}$. Then the coarse law is Av_b and its ORC is $A\Sigma A^\top$. Nonzero contrasts between fine categories can lie in the null space of A. For example, suppose a fair block chooses between deterministic categories z_1, z_2, both scored incorrect. Fine-category ORC has the nonzero submatrix

$$\begin{pmatrix} 1/4 & -1/4 \\ -1/4 & 1/4 \end{pmatrix},$$

whereas the correctness score is constant. The SQL example illustrates how coarsening the readout merges these distinctions.

For a general feature vector X, zero ORC means that its conditional mean is constant. A one-hot encoding of a finite path-label tuple $Z = (Z_{s_1}, \ldots, Z_{s_L})$ specifies the complete law of that tuple. Encoding the tuple thus captures information about the block carried by dependencies among future observations; Appendix C.2 gives an exact construction.

A.2 Variance budget and agreement between continuations

Conditional total covariance gives

$$C = \mathrm{Cov}(X) = \Sigma + \mathbb{E}\,\mathrm{Cov}(X \mid B).$$

For one-hot X, $C = \mathrm{Diag}(q) - qq^\top$. Projection onto a scalar contrast shows $0 \le \lambda_R \le \mathrm{Var}(R)$, establishing the range of ρ_R in (3). For a panel distribution over tasks, averaging both terms preserves this inequality; the reported ratio of averages is $\mathbb{E}_x \lambda_R(x) / \mathbb{E}_x \mathrm{Var}(R \mid x)$.

$$\Pr_{\text{independent}}(Z \neq Z') - \Pr_{\text{shared}}(Z \neq Z') = \operatorname{tr}\Sigma. \tag{12}$$

The first design samples independent blocks; the second shares one sampled block. Both use independent suffixes conditional on their blocks. Two futures with independently sampled blocks agree with probability $\|q\|^2$. Two conditionally independent futures sharing a block agree with probability $\mathbb{E}\|v_B\|^2$. Subtracting the first agreement probability from the second gives $\operatorname{tr}\Sigma$, which proves (12). With a Bernoulli readout of mean q_R, the two disagreement probabilities are $2q_R(1-q_R)$ and $2\{q_R(1-q_R) - \lambda_R\}$.

Proposition A.2 (Concentrated choice and normalized resolution)**.** *For $R \in [0,1]$, if some block has probability at least $1-\epsilon$, where $0 \leq \epsilon \leq 1/2$, then $\lambda_R \leq \epsilon(1-\epsilon)$. Two deterministic outcomes under block probabilities $1-\epsilon, \epsilon$ attain the bound. For $\epsilon > 0$, they have $\rho_R = 1$ even as $\lambda_R \to 0$.*

Proof. Write the block weights as w_b and their conditional rewards as $r_b \in [0,1]$. The quadratic form $r^\top(\operatorname{Diag}(w) - ww^\top)r$ is convex, so its maximum on the cube occurs at a vertex. Its value at a vertex is $u(1-u)$, where u is the weight assigned reward one. A block of mass at least $1-\epsilon$ implies either $u \leq \epsilon$ or $u \geq 1-\epsilon$. The bound and construction follow. □

Sampling entropy, normalized resolution, and absolute response capacity describe distinct properties. A nearly deterministic choice can resolve the full available outcome variance while having arbitrarily small absolute capacity. Conversely, many equiprobable choices can share the same future law and have zero capacity.

A.3 Sequential outcome budget

Let $M_s^X = \mathbb{E}_p[X \mid \mathcal{F}_s]$ and $D_s = M_{s+1}^X - M_s^X$, where X is a single bounded $\mathcal{F}_T$-measurable vector. The block covariance is $\Sigma_{t,k}^X = \operatorname{Cov}(M_{t+k}^X \mid \mathcal{F}_t)$. The vector martingale increments satisfy

$$\mathbb{E}[D_s \mid \mathcal{F}_s] = 0, \qquad \Sigma_s^X = \mathbb{E}[D_s D_s^\top \mid \mathcal{F}_s].$$

For $r < s$, conditioning first on $\mathcal{F}_s$ gives $\mathbb{E}[D_r D_s^\top \mid \mathcal{F}_t] = 0$; its transpose also vanishes. Consequently,

Proposition A.3 (Resolution budget for a single process)**.** *For a fixed target X under one process p,*

$$\Sigma_{t,k}^X = \operatorname{Cov}(M_{t+k}^X \mid \mathcal{F}_t) = \mathbb{E}\left[\sum_{s=t}^{t+k-1} \Sigma_s^X \,\middle|\, \mathcal{F}_t\right], \tag{13}$$

$$\operatorname{Cov}(X \mid \mathcal{F}_t) = \mathbb{E}\left[\sum_{s=t}^{T-1} \Sigma_s^X \,\middle|\, \mathcal{F}_t\right]. \tag{14}$$

A bounded stopping time $\tau \in [t,T]$ replaces the first sum by $\sum_{s=t}^{T-1} \mathbf{1}\{s < \tau\}\Sigma_s^X$ and its left side by $\operatorname{Cov}(M_\tau^X \mid \mathcal{F}_t)$.

Proof. Expand the covariance of $\sum_{s=t}^{t+k-1} D_s$ and apply the orthogonality above. For the stopped version, $\mathbf{1}\{s < \tau\}$ is $\mathcal{F}_s$-measurable, so stopped increments are still martingale differences. Their sum is $M_\tau^X - M_t^X$. At the full horizon, $M_T^X = X$. □

This gives the positive-semidefinite ordering of nested windows for one fixed target. At the token level, if $R = c_Z$, $q_t = \mathbb{E}[\boldsymbol{e}_Z \mid \mathcal{F}_t]$, and $\nu_{t,a}^p = \mathbb{E}_p[\boldsymbol{e}_Z \mid H_t, Y_{t+1} = a]$ on the policy support, then

$$\lambda_t(c) = \mathbb{E}[(c^\top q_{t+1} - c^\top q_t)^2 \mid \mathcal{F}_t].$$

The budget averages these contributions over all reachable histories. This conditional-variance accounting connects the budget to return-variance recursions (Sobel, 1982; Tamar et al., 2013; 2016).

B Information Geometry of Outcome Laws

Normalizing ORC by outcome base rates relates it to dependence measures and gives an information-geometric interpretation of the resolution fraction in Equation (3).

B.1 Information geometry and canonical outcome directions

In this subsection, denote the sampled block by A and write $\boldsymbol{\nu}_A = v_A$. Fix a prefix and abbreviate $A \sim \mu$, $\boldsymbol{q} = \mathbb{E}_\mu \boldsymbol{\nu}_A$, $\boldsymbol{\Sigma} = \mathrm{Cov}_\mu(\boldsymbol{\nu}_A)$, $\boldsymbol{D} = \mathrm{Diag}(\boldsymbol{q})$, and $S = \mathrm{supp}(\boldsymbol{q})$. Define

$$\mathcal{R}_\mu g := \mathbb{E}_\mu[g(A)(\boldsymbol{\nu}_A - \boldsymbol{q})], \qquad g \in L_0^2(\mu). \tag{15}$$

Its range lies in the outcome-simplex tangent space. On the supported outcome coordinates, equip this range with $\langle u, v\rangle_{\boldsymbol{D}^{-1}} := u^\top \boldsymbol{D}^+ v$. This normalization measures dependence relative to outcome base rates.

Theorem B.1 (Classical χ^2 and canonical-correlation identities)**.** *Let $Z \mid A = a \sim \boldsymbol{\nu}_a$. Then*

$$I_{\chi^2}(A; Z) := \chi^2(P_{A,Z} \| P_A \otimes P_Z) = \mathrm{tr}(\boldsymbol{D}^+ \boldsymbol{\Sigma}). \tag{16}$$

The nonzero eigenvalues of $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{D}_S^{-1/2} \boldsymbol{\Sigma}_{SS} \boldsymbol{D}_S^{-1/2}$ are the squared canonical correlations between functions of A and functions of Z; when both marginals are nondegenerate, its largest eigenvalue is the squared Hirschfeld–Gebelein–Rényi maximal correlation. Equivalently, the singular values of $\mathcal{R}_\mu$ in the $\boldsymbol{D}^{-1}$ outcome metric are those canonical correlations. Moreover,

$$0 \preceq \widetilde{\boldsymbol{\Sigma}} \preceq I - \sqrt{\boldsymbol{q}_S}\sqrt{\boldsymbol{q}_S}^\top, \quad \mathrm{rank}(\boldsymbol{\Sigma}) \le \min\{|S| - 1, |\operatorname{supp} \mu| - 1\}. \tag{17}$$

For binary Z with $0 < q < 1$, $I_{\chi^2}(A; Z) = \lambda/[q(1-q)] = \rho$. Operator-norm and path-information identities follow below.

B.2 Additional information identities

With $\widetilde{\boldsymbol{\Sigma}}$ as in Theorem B.1,

$$\|\mathcal{R}_\mu\|^2_{\mathrm{op}, \boldsymbol{D}^{-1}} = \lambda_{\max}(\widetilde{\boldsymbol{\Sigma}}), \qquad \|\mathcal{R}_\mu\|^2_{\mathrm{HS}, \boldsymbol{D}^{-1}} = \mathrm{tr}(\boldsymbol{D}^+ \boldsymbol{\Sigma}) = I_{\chi^2}(A; Z). \tag{18}$$

The Shannon information satisfies

$$I(A; Z) \le \log(1 + I_{\chi^2}(A; Z)) \le I_{\chi^2}(A; Z). \tag{19}$$

For the autoregressive process and $\boldsymbol{D}_t = \mathrm{Diag}(\boldsymbol{q}_t)$,

$$\sum_{t=0}^{T-1} I(Z; Y_{t+1} \mid H_t) = H(Z \mid x), \qquad \sum_t \mathbb{E}\,\mathrm{tr}(\boldsymbol{D}_t^+ \boldsymbol{\Sigma}_t) \ge H(Z \mid x). \tag{20}$$

B.3 Chi-square information geometry

If $q_z = 0$, then $\nu_a(z) = 0$ for μ-almost every a. On S,

$$\begin{aligned}\chi^2(P_{A,Z} \| P_A \otimes P_Z) &= \sum_{a,z} \mu(a) q_z \left(\frac{\nu_a(z)}{q_z} - 1\right)^2 \\ &= \sum_{z \in S} \frac{1}{q_z} \sum_a \mu(a)(\nu_a(z) - q_z)^2 = \mathrm{tr}(\boldsymbol{D}^+ \boldsymbol{\Sigma}).\end{aligned}$$

Let $N_{az} = \sqrt{\mu(a)}(\nu_a(z) - q_z)/\sqrt{q_z}$. Then $N^\top N = \widetilde{\boldsymbol{\Sigma}}$. This is the centered standardized joint-law matrix, whose singular values are the canonical correlations of the two finite random variables and whose largest singular value is HGR maximal correlation (Lancaster, 1958; Rényi, 1959; Buja, 1990). Total covariance gives

$$\mathrm{Diag}(\boldsymbol{q}) - \boldsymbol{q}\boldsymbol{q}^\top = \boldsymbol{\Sigma} + \mathbb{E}_\mu[\mathrm{Diag}(\boldsymbol{\nu}_A) - \boldsymbol{\nu}_A \boldsymbol{\nu}_A^\top] \succeq \boldsymbol{\Sigma}.$$

Conjugating by $\boldsymbol{D}_S^{-1/2}$ proves the Loewner bound. The rank bound follows because centered candidate laws span at most $|\operatorname{supp} \mu| - 1$ dimensions and all lie in the $|S| - 1$ dimensional simplex tangent space.

Equip the outcome space with $\langle u, v\rangle_{\boldsymbol{D}^{-1}} = u^\top \boldsymbol{D}^+ v$. The adjoint of the response map in this metric is $(\mathcal{R}_\mu^\dagger c)(a) = (\boldsymbol{\nu}_a - \boldsymbol{q})^\top \boldsymbol{D}^+ c$; hence $\mathcal{R}_\mu \mathcal{R}_\mu^\dagger = \boldsymbol{\Sigma}\boldsymbol{D}^+$, which is similar on S to $\widetilde{\boldsymbol{\Sigma}}$. Its eigenvalues are

therefore the squared singular values of $\mathcal{R}_\mu$, proving $\|\mathcal{R}_\mu\|^2_{\mathrm{op},\boldsymbol{D}^{-1}} = \lambda_{\max}(\widetilde{\boldsymbol{\Sigma}})$ and $\|\mathcal{R}_\mu\|^2_{\mathrm{HS},\boldsymbol{D}^{-1}} = \mathrm{tr}(\boldsymbol{D}^{+}\boldsymbol{\Sigma}) = I_{\chi^2}(A;Z)$. In the binary case, $\boldsymbol{\Sigma} = \lambda \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, so $\mathrm{tr}(\boldsymbol{D}^{-1}\boldsymbol{\Sigma}) = \lambda/[q(1-q)]$.

For any $P \ll Q$, Jensen's inequality gives $D_{\mathrm{KL}}(P\|Q) \le \log(1+\chi^2(P\|Q)) \le \chi^2(P\|Q)$. The chain rule and the fact that Z is determined by the completed path give

$$\sum_t I(Z; Y_{t+1} \mid H_t) = I(Z; Y_{1:T} \mid x) = H(Z \mid x).$$

Applying the one-step inequality conditionally and averaging yields (20).

C Sequential Cost, Temporal Alignment, and Local Response

C.1 Sequential factorization and continuation cost

Every supported finite block law $\nu_B \ll \mu_B$ admits autoregressive conditionals $\nu_j(b_j \mid h, b_{<j})$. When μ_B is the natural block law, likelihood-ratio factorization gives

Proposition C.1 (Sequential cost)**.**

$$D_{\mathrm{KL}}(\nu_B\|\mu_B) = \sum_j \mathbb{E}_\nu D_{\mathrm{KL}}\big(\nu_j(\cdot \mid h, B_{<j})\|p(\cdot \mid h, B_{<j})\big), \tag{21}$$

$$D_{\mathrm{KL}}(\nu_B K\|\mu_B K) = D_{\mathrm{KL}}(\nu_B\|\mu_B), \tag{22}$$

$$D_{\mathrm{KL}}(\mathcal{L}_{\nu_B K}(X)\|\mathcal{L}_{\mu_B K}(X)) \le D_{\mathrm{KL}}(\nu_B\|\mu_B). \tag{23}$$

Here the joint laws in the middle line retain both block and suffix; X in the final line is any common measurable readout.

Proof. The block log likelihood ratio is the sum of its sequential log ratios. Taking conditional expectations proves the first identity, including the zero contribution of absorbing EOS padding. The common suffix kernel cancels in the joint likelihood ratio, which proves the second identity. Marginalization to a readout gives the third by the log-sum inequality. □

Total KL therefore provides a common perturbation budget for blocks of different lengths. A forced supported block b has cost $-\log \mu_B(b)$, with its later conditionals evaluated on the changed history. A reference law supported on additional alternatives defines its own finite comparison process.

C.2 Joint checkpoint responses

Fix checkpoint maps X_s and their conditional means $m_s(B)$. Checkpoints may be absolute token positions, or events with a common bounded stopping and scoring rule; in both cases the readout is a bounded function of the completed trace. Let V stack these means. Then

$$a^\top \Gamma a = \mathbb{E}_\mu[a^\top (V - \mathbb{E}V)]^2 \ge 0.$$

Consequently, $\Gamma_{s,s} = \Sigma^{(s)}$, $\Gamma_{r,s} = \Gamma_{s,r}^\top$. For scalar readouts with fixed score orientations, each cross-entry satisfies

$$|\Gamma_{s,r}| \le \sqrt{\lambda_s \lambda_r}.$$

The off-diagonal sign measures the alignment of conditional responses to the same block.

The examples in the main text can be implemented by an autoregressive process that first reveals independent fair bits (U, V) and subsequently emits their selected copies or complements. The three pairs (U, U), (U, V), and $(U, 1-U)$ have covariance matrices with equal diagonals and cross-entries $1/4, 0, -1/4$. For a further independent suffix bit W, the sequence of readouts (U, W, U) has

$$\Gamma = \begin{pmatrix} 1/4 & 0 & 1/4 \\ 0 & 0 & 0 \\ 1/4 & 0 & 1/4 \end{pmatrix}.$$

Its first and last readouts capture the same contrast, while the middle readout is insensitive to it. The budget for increasing windows concerns a fixed target; this construction changes the target with observation time.

Marginal response and complete joint laws. Let the block reveal a fair bit U and let a fresh fair suffix bit be W. Put $X_s = W$ and $X_r = W \oplus U$. Both conditional marginal means are $1/2$ under either block, so the stacked marginal response matrix is zero. In contrast, the conditional tuple laws, in order $(00, 01, 10, 11)$, are

$$(1/2, 0, 0, 1/2) \quad \text{and} \quad (0, 1/2, 1/2, 0).$$

Their one-hot tuple ORC has trace $1/4$. Equivalently, the event $\mathbf{1}\{X_s = X_r\}$ has capacity $1/4$. Stacked checkpoint means thus describe temporal alignment of the chosen observables, while encoding the tuple also captures how their relationships depend on the block.

C.3 Response operator and directional alignment

Let a centered score $g(B)$ define a local reweighting $\mu_\varepsilon(b) \propto \mu(b)e^{\varepsilon g(b)}$, retaining K. Its squared score norm is the local Fisher cost, since $D_{\mathrm{KL}}(\mu_\varepsilon \| \mu) = \varepsilon^2 \mathbb{E}_\mu g^2/2 + o(\varepsilon^2)$. Stack the checkpoint means as $V(B) = (m_{s_1}(B)^\top, \ldots, m_{s_L}(B)^\top)^\top$. The response operator satisfies

$$\mathcal{T} g = \frac{d}{d\varepsilon} \mathbb{E}_{\mu_\varepsilon K}(X_{s_1}^\top, \ldots, X_{s_L}^\top)^\top \Big|_0 = \mathbb{E}_\mu[(V - \mathbb{E}_\mu V)g], \qquad \mathcal{T}\mathcal{T}^* = \Gamma. \tag{24}$$

This applies the score identity and Fisher geometry (Williams, 1992; Kakade, 2001) jointly to the specified checkpoint readouts.

Proposition C.2 (Outcome-specific response)**.** *For a scalar target R with $m_R(b) = \mathbb{E}_K[R \mid h, b]$, put $\lambda_R = \mathrm{Var}_\mu\, m_R(B)$. The first-order response is*

$$\frac{d}{d\varepsilon} \mathbb{E}_{\mu_\varepsilon K} R \Big|_0 = \|g\|_{L^2(\mu)} \sqrt{\lambda_R}\, \alpha_R(g), \qquad |\alpha_R(g)| \le 1, \tag{25}$$

where $\alpha_R(g)$ is the correlation between g and $m_R(B)$, set to zero when either variance vanishes. Under $\mathbb{E}g^2 \le 1$, the maximum response is $\sqrt{\lambda_R}$, attained by the normalized centered m_R when $\lambda_R > 0$.

Proof of Proposition C.2 and (24)*.* On the finite block support, let $\mathcal{H} = L_0^2(\mu)$ with its probability-weighted inner product. Differentiating the normalized exponential tilt yields

$$\mathcal{T} g = \mathbb{E}_\mu[(V - \mathbb{E}V)g], \qquad \mathcal{T}^* a = a^\top (V - \mathbb{E}V).$$

Hence $\mathcal{T}\mathcal{T}^* a = \Gamma a$. A scalar projection R has derivative $\langle g, m_R - \mathbb{E}m_R \rangle_{L^2(\mu)}$. Cauchy–Schwarz gives (25) and the unit-norm maximum. The maximizing score is $(m_R - \mathbb{E}m_R)/\sqrt{\lambda_R}$ when $\lambda_R > 0$; at zero capacity every supported centered score has zero derivative. □

A direction learned for one score has its own inner product with every other score's conditional mean. These inner products determine the first-order target responses and explain target alignment in the branch-reweighting study.

C.4 Finite reweighting and gains under a small KL budget

For $\nu \ll \mu$, write $w = \nu/\mu$, $q_R = \mathbb{E}_\mu m_R$, and $\lambda_R = \mathrm{Var}_\mu\, m_R$. Then

$$\mathbb{E}_{\nu K} R - \mathbb{E}_{\mu K} R = \mathbb{E}_\mu[(w - 1)(m_R - q_R)].$$

Since $\mathbb{E}_\mu(w - 1) = 0$ and $\mathbb{E}_\mu(w - 1)^2 = \chi^2(\nu \| \mu)$, normalizing this inner product gives the exact capacity–alignment factorization in (5); Cauchy–Schwarz gives its bound. On a finite support, define

$$G_\delta(h; R, K) = \sup_{\nu : D_{\mathrm{KL}}(\nu \| \mu) \le \delta} \{\mathbb{E}_{\nu K} R - \mathbb{E}_{\mu K} R\}. \tag{26}$$

For $\lambda_R > 0$,

$$G_\delta(h; R, K) = \sqrt{2\delta\lambda_R} + O(\delta), \qquad \delta \downarrow 0. \tag{27}$$

To establish the lower and upper bounds, let $Z = m_R - q_R$ and $\psi(\eta) = \log \mathbb{E}_\mu e^{\eta Z}$. Finite support gives $\psi(\eta) = \eta^2 \lambda_R/2 + O(\eta^3)$. The positive exponential tilt has gain $\psi'(\eta) = \eta\lambda_R + O(\eta^2)$ and KL $\eta\psi'(\eta) - \psi(\eta) = \eta^2\lambda_R/2 + O(\eta^3)$, attaining the expansion. For any feasible ν and $\eta > 0$, nonnegativity of relative entropy to this tilt gives $\mathbb{E}_\nu Z \le \{\delta + \psi(\eta)\}/\eta$. Taking $\eta = \sqrt{2\delta/\lambda_R}$ gives the matching upper bound. The optimization here is over all distributions in the supported KL ball; a restricted score family defines its own achievable subset.

D Position-Constrained Response Along a Generation

This section extends the fixed-prefix response in Equation (5) to permitted intervention positions along a generation, showing how capacities at different positions combine under a Fisher budget for the full path.

For this section, $q_t = \mathbb{E}_p[\boldsymbol{e}_Z \mid \mathcal{F}_t]$ and $\nu^p_{t,a}$ is the token-conditioned categorical future law under the baseline process p. Let $c_Z \in [0,1]$. A differentiable policy perturbation has bounded conditionally centered scores $s_t(a,h) = \partial_\epsilon \log p_{\epsilon,t}(a \mid h)|_0$. For a predictable mask $\omega_t \in \{0,1\}$, define

$$J'_c(0) = \sum_t \mathbb{E}[s_t(Y_{t+1}, H_t) c^\top (q_{t+1} - q_t)], \qquad \mathcal{K}_\omega(c) = \mathbb{E}\sum_t \omega_t \lambda_t(c). \tag{28}$$

The mask identifies permitted intervention positions before their next token is sampled.

Proposition D.1 (Exact first-order capacity on selected positions)**.** *Among bounded conditionally centered scores supported on ω and satisfying the path-Fisher budget $\mathbb{E}\sum_t s_t^2 \le B$,*

$$\sup |J'_c(0)| = \sqrt{B\mathcal{K}_\omega(c)}. \tag{29}$$

For $\mathcal{K}_\omega(c) > 0$, the score

$$s^*_t(a,h) = \sqrt{B/\mathcal{K}_\omega(c)}\, \omega_t(h) c^\top \{\nu^p_{t,a}(h) - q_t(h)\} \tag{30}$$

attains equality.

D.1 Pathwise score identity and exact capacity

Differentiating the finite path sum gives

$$J'_{\boldsymbol{c}}(0) = \mathbb{E}\left[c_Z \sum_t s_t(Y_{t+1}, H_t)\right].$$

Since s_t is $\mathcal{F}_{t+1}$-measurable and centered given $\mathcal{F}_t$,

$$\mathbb{E}[s_t c_Z] = \mathbb{E}[s_t \mathbb{E}(c_Z \mid \mathcal{F}_{t+1})] = \mathbb{E}[s_t \boldsymbol{c}^\top (\boldsymbol{q}_{t+1} - \boldsymbol{q}_t)],$$

which proves (28). Conditional Cauchy–Schwarz then gives

$$|J'_{\boldsymbol{c}}(0)| \le \left(\mathbb{E}\sum_t s_t^2\right)^{1/2} \left(\mathbb{E}\sum_t \omega_t [\boldsymbol{c}^\top (\boldsymbol{q}_{t+1} - \boldsymbol{q}_t)]^2\right)^{1/2} \le \sqrt{B\mathcal{K}_\omega(\boldsymbol{c})}.$$

The field (30) is centered, supported on ω, has squared norm B, and attains equality.

D.2 Natural-gradient response and parameter sharing

Write $a_t(a,h) = c^\top\{\nu^p_{t,a}(h) - q_t(h)\}$ for the conditional advantage. In the unrestricted conditional-score space,

$$\|\nabla_F J_{\boldsymbol{c}}\|_F^2 = \mathbb{E}\sum_t a_t(Y_{t+1})^2 = \mathbb{E}\sum_t \lambda_t(\boldsymbol{c}) = \mathrm{Var}(c_Z \mid x). \tag{31}$$

Use the inner product $\langle s, r\rangle_F = \mathbb{E}\sum_t s_t r_t$ on conditionally centered score fields. The path derivative has Riesz representer $a_t(Y_{t+1})$, so its squared dual norm is $\mathbb{E}\sum_t \lambda_t = \mathrm{Var}(c_Z \mid x)$. Restricting to a differentiable shared-parameter family projects this representer onto its score span. If F_θ is that family's path Fisher matrix and $\nabla_\theta J$ its gradient, then

$$(\nabla_\theta J)^\top F_\theta^+ (\nabla_\theta J) \le \mathbb{E}\sum_t \lambda_t,$$

with equality precisely when the advantage field belongs to the attainable score span. This projection identifies the portion of outcome response accessible to a shared-parameter update.

D.3 CAPACITY-OPTIMAL PLACEMENT

Proposition D.2 (Capacity-optimal placement). *For deterministic sets $S \subseteq \{0, \dots, T-1\}$ with $|S| \le k$, $\sum_{t\in S} \mathbb{E}\lambda_t(\boldsymbol{c})$ is maximized by the k largest values of $\mathbb{E}\lambda_t(\boldsymbol{c})$. For $0 \le \kappa \le T$, allow independent auxiliary randomization. Among predictable rules $\omega_t \in \{0,1\}$ satisfying $\mathbb{E}\sum_t \omega_t \le \kappa$, an optimal rule thresholds the current capacity,*

$$\omega_t^\tau = \mathbf{1}\{\lambda_t(\boldsymbol{c}) > \tau\}, \tag{32}$$

with randomization at equality when needed to meet the budget. Combining with Proposition D.1, this rule maximizes first-order control capacity under the expected position budget.

For static sets, additivity makes the top-k rule immediate. For the expected-count problem, choose τ and boundary randomization so that $\mathbb{E}\sum_t \omega_t^\tau = \kappa$. Pointwise in (t, H_t), ω_t^τ maximizes $\omega(\lambda_t - \tau)$ over $\omega \in [0,1]$. Therefore, for every feasible predictable ω,

$$\mathcal{K}_\omega - \tau\mathbb{E}\sum_t \omega_t \le \mathcal{K}_{\omega^\tau} - \tau\kappa.$$

Since $\mathbb{E}\sum_t \omega_t \le \kappa$ and $\tau \ge 0$, $\mathcal{K}_\omega \le \mathcal{K}_{\omega^\tau}$. The capacity claim follows from Proposition D.1. The budget constrains the expected number of selected positions, allowing that number to vary between realized paths.

D.4 FINITE EXPONENTIAL TILTS

The following guarantee extends the first-order capacity result to finite tilts along the capacity direction. Set

$$g_t(a,h) = \omega_t(h)\,\boldsymbol{c}^\top(\boldsymbol{\nu}^p_{t,a}(h) - \boldsymbol{q}_t(h)), \qquad p_{\varepsilon,t}(a \mid h) \propto p_t(a \mid h)e^{\varepsilon g_t(a,h)}, \tag{33}$$

assume $c_Z \in [0,1]$, and suppose $\sum_t \omega_t \le k$ on every path. The score field and predictable selection rule are computed under the baseline policy and held fixed as ε varies.

Theorem D.3 (Finite-tilt response and KL cost). *Let $J(\varepsilon) = \mathbb{E}_\varepsilon[c_Z]$ and $K_\omega = \mathcal{K}_\omega(\boldsymbol{c})$. For every $\varepsilon \in \mathbb{R}$,*

$$\begin{aligned} |J(\varepsilon) - J(0) - \varepsilon K_\omega| &\le k\varepsilon^2/8, \\ D_{\mathrm{KL}}(P_\varepsilon \| P) &\le k\varepsilon^2/8. \end{aligned} \tag{34}$$

If $K_\omega > 0$, the explicit choice $\varepsilon_* = 4K_\omega/k$ gives

$$J(\varepsilon_*) - J(0) \ge \frac{2K_\omega^2}{k}, \qquad D_{\mathrm{KL}}(P_{\varepsilon_*}\|P) \le \frac{2K_\omega^2}{k}. \tag{35}$$

At fixed capacity, fewer active positions yield a larger guaranteed finite gain. At one prefix, the exact advantage tilt increases expected reward monotonically.

The finite-tilt bound follows from the next bounded-score envelope.

Proposition D.4 (General bounded-score finite-tilt envelope). *Let $g_t(\cdot, h)$ be centered under $p_t(\cdot \mid h)$, vanish whenever a predictable mask $\omega_t(h)$ is zero, and satisfy $|g_t(a,h)| \le \bar{G}$. Define*

$$p_{\varepsilon,t}(a \mid h) = \frac{p_t(a \mid h)e^{\varepsilon g_t(a,h)}}{\sum_b p_t(b \mid h)e^{\varepsilon g_t(b,h)}}, \qquad J(\varepsilon) = \mathbb{E}_\varepsilon[c_Z], \quad c_Z \in [0,1],$$

and let

$$I(\varepsilon) = \mathbb{E}_\varepsilon \sum_t \mathrm{Var}_{p_{\varepsilon,t}(\cdot|H_t)}(g_t), \quad B_\infty = \operatorname*{ess\,sup}_P \sum_t \mathrm{Var}_{p_t(\cdot|H_t)}(g_t), \quad N_\infty = \operatorname*{ess\,sup}_P \sum_t \omega_t.$$

Then J is smooth,

$$J'(0) = \sum_t \mathbb{E}[g_t(Y_{t+1}, H_t)(M_{t+1} - M_t)], \qquad M_t = \mathbb{E}[c_Z \mid \mathcal{F}_t], \tag{36}$$

$|J''(\varepsilon)| \le I(\varepsilon)$, and

$$\begin{aligned} |J(\varepsilon) - J(0) - \varepsilon J'(0)| &\le \frac{\varepsilon^2}{2}\,\overline{I}(\varepsilon), \\ D_{\mathrm{KL}}(P_\varepsilon\|P) &\le \frac{\varepsilon^2}{2}\,\overline{I}(\varepsilon), \end{aligned} \qquad \overline{I}(\varepsilon) := \min\left\{e^{2|\varepsilon|\bar{G}}B_\infty, \bar{G}^2 N_\infty\right\}. \tag{37}$$

Proof. Write

$$\ell_\varepsilon = \log \frac{dP_\varepsilon}{dP} = \sum_t \{\varepsilon g_t - \psi_t(\varepsilon, H_t)\}, \qquad \psi_t(\varepsilon, h) = \log \mathbb{E}_{p_t(\cdot|h)} e^{\varepsilon g_t}.$$

Let $\widetilde{g}_{\varepsilon,t} = g_t - \psi_t'(\varepsilon, H_t)$. Under P_ε, these are martingale differences and

$$\partial_\varepsilon \ell_\varepsilon = \sum_t \widetilde{g}_{\varepsilon,t}, \qquad \partial_\varepsilon^2 \ell_\varepsilon = -\sum_t \mathrm{Var}_{p_{\varepsilon,t}}(g_t).$$

Differentiation of the finite path sum gives $J'(\varepsilon) = \mathbb{E}_\varepsilon[c_Z \sum_t \widetilde{g}_{\varepsilon,t}]$ and

$$J''(\varepsilon) = \mathbb{E}_\varepsilon \left[c_Z \left(\sum_t \widetilde{g}_{\varepsilon,t} \right)^2 \right] - \mathbb{E}_\varepsilon \left[c_Z \sum_t \mathrm{Var}_{p_{\varepsilon,t}}(g_t) \right].$$

Orthogonality of martingale differences gives $\mathbb{E}_\varepsilon(\sum_t \widetilde{g}_{\varepsilon,t})^2 = I(\varepsilon)$. Since $c_Z \in [0,1]$, both terms in the preceding display lie in $[0, I(\varepsilon)]$, proving $|J''(\varepsilon)| \le I(\varepsilon)$. At $\varepsilon = 0$, conditioning successively on $\mathcal{F}_t$ and $\mathcal{F}_{t+1}$ yields (36).

For any centered g with $|g| \le \bar{G}$,

$$\mathrm{Var}_{p_\varepsilon}(g) \le \min\{e^{2|\varepsilon|\bar{G}} \mathrm{Var}_p(g), \bar{G}^2\}.$$

Summing this pointwise inequality along a path gives $I(u) \le \overline{I}(\varepsilon)$ for $|u| \le |\varepsilon|$. Taylor's integral remainder proves the first bound in (37). All finite tilts have the same path support as P. Apply the conditional KL identity $D_{\mathrm{KL}}(p_\varepsilon \| p) = \int_0^\varepsilon u \, \mathrm{Var}_{p_u}(g)\, du$ at each prefix, then average its path sum under P_ε. The same pointwise envelope and the KL chain rule prove the second bound. □

Specialization to the capacity direction. For (33), at every prefix the score values lie in an interval of length at most $\omega_t(h)$, because $c^\top \nu_{t,a}^p \in [0,1]$. Therefore, for every $u \in \mathbb{R}$, $\mathrm{Var}_{p_{u,t}(\cdot|h)}(g_t) \le \omega_t(h)/4$. Since at most k positions are active on each path, $I(u) \le k/4$. Taylor's integral remainder and the conditional KL integral in the preceding proof give (34). Moreover, $J'(0) = K_\omega$, and substitution of $\varepsilon_* = 4K_\omega/k$ gives (35). This sharpens the general absolute-score envelope by using the score's oscillation.

E Paired Estimation Details

E.1 Equivalent noise-corrected and U-statistic forms

Write $\widehat{\Sigma} = \widehat{\Sigma}_U$ in this appendix. For $m, n \ge 2$, let

$$W_i = \frac{1}{n(n-1)} \sum_{r \ne s} X_{ir} X_{is}^\top.$$

Expanding the sample covariance within each block gives $W_i = \bar{X}_i \bar{X}_i^\top - S_i/n$, hence

$$\widehat{\Sigma}_U = \frac{1}{m} \sum_i W_i - \frac{1}{m(m-1)} \sum_{i \ne j} \bar{X}_i \bar{X}_j^\top. \tag{38}$$

This is algebraically equivalent to the noise-corrected form in (6).

Proposition E.1 (Paired identification)**.** *If blocks are iid and suffix replicas are conditionally iid given each block,* (38) *is unbiased for* $\Sigma = \mathrm{Cov}_\mu \mathbb{E}[X \mid B]$ *for every* $m, n \ge 2$*. The statement applies to a bounded feature vector as well as a one-hot label.*

Proof. Conditional independence gives $\mathbb{E}[W_i \mid B_i] = v_{B_i} v_{B_i}^\top$. Independence of blocks gives $\mathbb{E}[\bar{X}_i \bar{X}_j^\top] = qq^\top$ for $i \ne j$. Subtraction yields the result. □

For $n = 2$, the symmetrized covariance of the suffix columns is

$$\widehat{\Sigma}_{\mathrm{col}} = \frac{1}{2(m-1)} \sum_i \left[(X_{i1} - \bar{X}_{\cdot 1})(X_{i2} - \bar{X}_{\cdot 2})^\top + (X_{i2} - \bar{X}_{\cdot 2})(X_{i1} - \bar{X}_{\cdot 1})^\top \right]. \tag{39}$$

Its checkpoint cross-blocks estimate $\Gamma_{s,r}$. With $d_i = X_{i1} - X_{i2}$, expansion gives the exact finite-sample relation

$$\widehat{\Sigma}_{\mathrm{col}} - \widehat{\Sigma}_U = \frac{(\sum_i d_i)(\sum_i d_i)^\top - \sum_i d_i d_i^\top}{4m(m-1)}. \tag{40}$$

Its expectation is zero because $\mathbb{E}d_i = 0$ and different blocks are independent. For scalar outcomes, $\widehat{\Sigma}_U = (\mathrm{MS}_B - \mathrm{MS}_W)/n$ exactly. Finite-sample estimates retain their signs; covariance identities refer to population quantities or the stated estimators.

For temporal observations, substitute the stacked checkpoint vector for X. A same-suffix early–late product has conditional expectation

$$m_s(B) m_r(B)^\top + \mathrm{Cov}_K(X_s, X_r \mid B),$$

whereas products from different suffix replicas have conditional expectation $m_s(B) m_r(B)^\top$. The latter expectation isolates the common-block component.

E.2 Exact variance and jackknife

Using the token notation, write $A = B$ and $\nu_A = v_B$ below. For any fixed matrix Θ, let $\theta = \langle \Theta, \mathbf{\Sigma}_t(\mu; p) \rangle$, $\Theta_s = (\Theta + \Theta^\top)/2$, and define

$$\zeta_1 = \mathrm{Var}(\bar{X}_1^\top \Theta_s \boldsymbol{q}), \qquad \zeta_2 = \mathrm{Var}(\bar{X}_1^\top \Theta_s \bar{X}_2).$$

Write $\widehat{\theta} = \langle \Theta, \widehat{\mathbf{\Sigma}} \rangle$, $\boldsymbol{q} = \mathbb{E}_\mu \boldsymbol{\nu}_A$, and

$$\xi_i = \langle \Theta, W_i \rangle - 2 \bar{X}_i^\top \Theta_s \boldsymbol{q}.$$

For fixed n and $\mathrm{Var}(\xi_1) > 0$,

$$\sqrt{m}(\widehat{\theta} - \theta) \Rightarrow \mathcal{N}(0, \mathrm{Var}(\xi_1)), \qquad m \to \infty. \tag{41}$$

Then the estimator in (6) satisfies the exact identity

$$\mathrm{Var}(\widehat{\theta}) = \frac{\mathrm{Var}(\xi_1)}{m} + \frac{2\zeta_2 - 4\zeta_1}{m(m-1)}. \tag{42}$$

Indeed, $\mathbb{E}W_i = \mathbb{E}_\mu[\boldsymbol{\nu}_A \boldsymbol{\nu}_A^\top]$ and, for $i \neq j$, $\mathbb{E}\bar{X}_i \bar{X}_j^\top = \boldsymbol{q}\boldsymbol{q}^\top$, proving unbiasedness. Write $\widehat{\theta} = \bar{w} - U$, where $\bar{w} = m^{-1} \sum_i \langle \Theta, W_i \rangle$ and $U = 2[m(m-1)]^{-1} \sum_{i<j} \bar{X}_i^\top \Theta_s \bar{X}_j$. Hoeffding's variance formula for a second-order U-statistic and direct covariance counting give (42). The Hoeffding decomposition yields

$$\widehat{\theta} - \theta = \frac{1}{m} \sum_{i=1}^{m} (\xi_i - \mathbb{E}\xi_i) + O_P(m^{-1}),$$

which proves (41). The delete-one-group jackknife is consistent for the leading variance by standard U-statistic theory (Arvesen, 1969) and can be computed by updating $\sum_i W_i$, $\sum_i \bar{X}_i$, and $\sum_i \bar{X}_i \bar{X}_i^\top$.

For $R_{ir} = \boldsymbol{c}^\top X_{ir}$, expanding the balanced mean squares gives

$$\frac{\mathrm{MS}_B - \mathrm{MS}_W}{n} = \frac{1}{m} \sum_i \frac{\sum_{r \neq s} R_{ir} R_{is}}{n(n-1)} - \frac{1}{m(m-1)} \sum_{i \neq j} \bar{R}_i \bar{R}_j,$$

which is exactly $\boldsymbol{c}^\top \widehat{\mathbf{\Sigma}} \boldsymbol{c}$.

E.3 WITHIN-TOKEN REPLICATION AND GROUP-SIZE CHOICE

With a fixed total of $N = 128$ suffixes, $n \in \{2, 4, 8, 16\}$ gives $(m, n) = (64, 2), (32, 4), (16, 8), (8, 16)$. The number of replicas n determines both within-token noise and the number $m = N/n$ of independently sampled tokens. The following decomposition quantifies both contributions to estimator variance.

Proposition E.2 (Group-size variance decomposition)**.** *Let X, X_1, X_2 be conditionally iid bounded vectors with conditional mean $\boldsymbol{\nu}_a$ given $A = a$, and define*

$$\zeta_1^w(a) = \mathrm{Var}(X^\top \Theta_s \boldsymbol{\nu}_a \mid A = a), \qquad \zeta_2^w(a) = \mathrm{Var}(X_1^\top \Theta_s X_2 \mid A = a).$$

For the influence variable $\xi_1 = \langle \Theta, W_1 \rangle - 2\bar{X}_1^\top \Theta_s \boldsymbol{q}$,

$$\begin{aligned} \mathrm{Var}(\xi_1) &= \mathrm{Var}\big(\boldsymbol{\nu}_A^\top \Theta \boldsymbol{\nu}_A - 2\boldsymbol{\nu}_A^\top \Theta_s \boldsymbol{q}\big) + \mathbb{E}\,\mathrm{Var}(\xi_1 \mid A), \\ \mathrm{Var}(\xi_1 \mid A = a) &= \frac{2\{2(n-2)\zeta_1^w(a) + \zeta_2^w(a)\}}{n(n-1)} \\ &\quad - \frac{8}{n}\,\mathrm{Cov}\big(X^\top \Theta_s \boldsymbol{\nu}_a, X^\top \Theta_s \boldsymbol{q} \mid A = a\big) + \frac{4}{n}\,\mathrm{Var}\big(X^\top \Theta_s \boldsymbol{q} \mid A = a\big) \\ &= \frac{4}{n}\,\mathrm{Var}\big(X^\top \Theta_s (\boldsymbol{\nu}_a - \boldsymbol{q}) \mid A = a\big) + O(n^{-2}). \end{aligned} \tag{43}$$

The first term in the unconditional decomposition is independent of n.

Proof. Condition on $A = a$ and write $\xi_1 = U_n^{(a)} - 2\bar{X}^\top \Theta_s \boldsymbol{q}$, where $U_n^{(a)}$ is the order-two U-statistic with kernel $h(x, y) = x^\top \Theta_s y$ and first projection $h_1^{(a)}(x) = x^\top \Theta_s \boldsymbol{\nu}_a$. Hoeffding's variance formula gives the first term in the conditional display. The sample-mean term has variance $4\,\mathrm{Var}(X^\top \Theta_s \boldsymbol{q} \mid a)/n$, while

$$\mathrm{Cov}\left(U_n^{(a)}, \frac{1}{n}\sum_{r=1}^{n} f(X_r) \middle| a\right) = \frac{2}{n}\,\mathrm{Cov}\Big(h_1^{(a)}(X), f(X) \mid a\Big).$$

Combining these identities gives the exact conditional formula; collecting its terms of order n^{-1} gives the final expansion. The unconditional statement follows from the law of total variance and $\mathbb{E}[\xi_1 \mid A = a] = \boldsymbol{\nu}_a^\top \Theta \boldsymbol{\nu}_a - 2\boldsymbol{\nu}_a^\top \Theta_s \boldsymbol{q}$. □

In $\mathrm{Var}(\xi_1)$, the between-token component is constant in n, and the leading within-token component decays as $1/n$. With prefix caching, a pilot study can choose a common $n \in \{2, 4, 8, 16\}$ by minimizing the product of estimated leading variance and cost per group, $\widehat{\mathrm{Var}}(\xi_1; n)(c_{\mathrm{group}} + n c_{\mathrm{suffix}})$. The terminal observation map is defined for every completion, assigning malformed or truncated completions their own outcome class and keeping every rollout in the balanced-group estimator.

E.4 PAIRED-GROUP LOWER BOUND

Let $A \in \{-1, +1\}$ be uniform and conditional suffix outcomes be iid Bernoulli with $q_a(\theta) = 1/2 + a\theta$, $\theta \in [1/4, 3/8]$. Trace ORC is $L(\theta) = 2\theta^2$. For $\theta_0 = 1/4$, $\theta_1 = 1/4 + 3\varepsilon$, and $0 < \varepsilon \le 1/24$, the trace gap exceeds 2ε. The Bernoulli KL bound gives, for a group with n suffixes,

$$D_{\mathrm{KL}}(P_{\theta_0}^{(n)} \| P_{\theta_1}^{(n)}) \le \frac{576}{7} n\varepsilon^2.$$

Indeed, the Bernoulli parameters differ by 3ε and the denominator $q_a(\theta_1)(1 - q_a(\theta_1))$ is at least $7/64$. For m independent groups the divergence is at most $(576/7)mn\varepsilon^2$. Le Cam's two-point argument and Pinsker's inequality imply $N = mn \ge c_0 \varepsilon^{-2}$ for uniform accuracy ε with probability at least $2/3$. For fixed n, this is $m = \Omega_n(\varepsilon^{-2})$. This establishes how the required number of suffixes scales for paired-group observations.

E.5 Response-targeted allocation after screening

Once the candidate tokens or groups have been specified, let group j have candidate probability π_j, score g_j, suffix reward variance σ_j^2, and n_j forced suffixes. Let $\mathcal{J}_+ = \{j : \pi_j|g_j|\sigma_j > 0\}$. The variance criterion for the response estimate is $\sum_{j\in\mathcal{J}_+} \pi_j^2 g_j^2 \sigma_j^2/n_j$. For the continuous relaxation, take $n_j \geq 0$, with $n_j > 0$ on $\mathcal{J}_+$ and $\sum_j n_j = N > 0$. If $\mathcal{J}_+ \neq \varnothing$, the Neyman allocation

$$n_j^* = N\frac{\pi_j|g_j|\sigma_j}{\sum_\ell \pi_\ell|g_\ell|\sigma_\ell} \tag{44}$$

minimizes this criterion, with minimum $(\sum_j \pi_j|g_j|\sigma_j)^2/N$ (Neyman, 1934). If $\mathcal{J}_+ = \varnothing$, the criterion is zero for every feasible allocation. For integer suffix counts, choose an allocation that meets the budget and any minimum sampling requirements. The allocation applies to a fixed response score and the stated suffix-cost budget.

E.6 Fixed candidate enumeration and the suffix-noise correction

Fixed candidate enumeration uses a declared candidate list and its probability weights. Denote these weights by w_a, with $\sum_a w_a = 1$. Candidate a has $n_a \geq 2$ independent scalar scores of mean m_a, variance σ_a^2, sample mean $\widehat{m}_a$, and unbiased sample variance s_a^2. The target is $\lambda = \sum_a w_a m_a^2 - (\sum_a w_a m_a)^2$. Independent candidate pools give

$$\mathbb{E}\left[\sum_a w_a\widehat{m}_a^2 - \left(\sum_a w_a\widehat{m}_a\right)^2\right] = \lambda + \sum_a w_a(1-w_a)\frac{\sigma_a^2}{n_a}, \tag{45}$$

$$\widehat{\lambda}_{\text{enum}} = \sum_a w_a\widehat{m}_a^2 - \left(\sum_a w_a\widehat{m}_a\right)^2 - \sum_a w_a(1-w_a)\frac{s_a^2}{n_a}. \tag{46}$$

The second expression is unbiased. If D, E are independent pools for the same frozen candidates, then

$$\widehat{\lambda}_{DE} = \sum_a w_a\widehat{m}_a^D\widehat{m}_a^E - \left(\sum_a w_a\widehat{m}_a^D\right)\left(\sum_a w_a\widehat{m}_a^E\right) \tag{47}$$

is a second unbiased reference. These formulas use the actual candidate weights and separate candidate enumeration from iid natural-block sampling. The experiment with independent pools uses this design to measure how much naive spread comes from finite suffix sampling.

E.7 Panel estimands and paired uncertainty

For tasks x with declared weights π_x, the panel matrix is $\Sigma_{\text{panel}} = \sum_x \pi_x\Sigma_x$, and the corresponding temporal matrix is $\Gamma_{\text{panel}} = \sum_x \pi_x\Gamma_x$. The total within-task variance is $\sum_x \pi_x \operatorname{Var}(R \mid x)$, distinct from the additional between-task component $\operatorname{Var}_x \mathbb{E}[R \mid x]$. Each empirical ratio uses a within-task denominator matched to its numerator.

Comparisons between models, interventions, or outcome readouts use the same resampled tasks. Database-level resampling keeps questions from the same database together. Independent suffixes identify the within-block covariance; independent block groups and task clusters determine the resampling units for each uncertainty calculation. Resampling complete checkpoint vectors preserves their joint dependence. Re-pairing complete vectors from the second replica gives a conditional reference distribution for testing independence between the two observed future vectors. The test evaluates a specified scalar or matrix statistic under this randomization law.

F Feedback Policy Value Across Process States

Fix the distribution of tasks x and natural blocks B. Let $a(x, B) \in \{0, 1\}$ be a gate computed from available source data and the frozen block. Ordinary continuation has mean score $m_R(x, B)$, while the specified feedback operation has mean $m_R^F(x, B)$. The operation includes its turn boundary,

feedback text, answer horizon, and stochastic continuation. Outside the gate, the policy uses ordinary continuation. Its value is

$$G_F = \mathbb{E}\big[a(x,B)\{m_R^F(x,B) - m_R(x,B)\}\big]. \tag{48}$$

Proposition F.1 (Process-state decomposition and available headroom)**.** *For a finite pretreatment category $D = D(x,B)$ and $R \in [0,1]$,*

$$G_F = \sum_d \Pr(a=1, D=d)\mathbb{E}[m_R^F - m_R \mid a=1, D=d], \tag{49}$$

$$G_F \le \mathbb{E}[a(x,B)\{1 - m_R(x,B)\}]. \tag{50}$$

Proof. Partition the expectation in (48) by D and a for the first identity. The second uses $m_R^F \le 1$ pointwise. □

A deterministic task calculator can define D by whether the recorded state would alter its computation. This partitions the policy gain into weighted contributions from distinct process states, preserving the total. Both the conditional improvement and the frequency of each state determine the aggregate value. The local-reweighting response in (25) holds the continuation policy K fixed, while G_F evaluates the fully specified feedback policy K_F at the same observed block.

G Representations and Behavioral Coordinates

Let $R_t(h)$ be a representation of histories and suppose a map Ψ_t^p satisfies

$$\Pr_p(Z = \cdot \mid H_t = h) = \Psi_t^p(R_t(h)). \tag{51}$$

Such a map exists precisely when histories with the same representation have the same categorical future law: if $R_t(h) = R_t(h')$, the two histories must have equal laws for Z. Necessity follows by applying Ψ_t^p. For sufficiency, assign the common law to each reachable representation; the collision condition makes that assignment well-defined. Both injective representations and representations whose merged histories share a future law satisfy this condition.

For candidate hidden states $F_t(a) = R_{t+1}(H_t \oplus a)$,

$$\Sigma_t(\mu; p) = \mathrm{Cov}_{A\sim\mu}\big(\Psi_{t+1}^p(F_t(A))\big). \tag{52}$$

ORC therefore compares candidate states through the outcome laws they induce. Injectivity results concern hidden representations (Nikolaou et al., 2026); the observable future-law map can send distinct representations to the same image. An invertible reparameterization of the hidden states preserves this covariance. The observable Z determines which distinctions remain in these behavioral coordinates; refining Z can expose structure that a coarser score merges.

H Empirical Methods and Additional Results

H.1 Panels, policies, and replication

The models are `Qwen/Qwen3.8-27B` and `google/gemma-4-31B-it`. Both run with vLLM 0.28.0 in `bfloat16` on an H100 80GB, with prefix caching disabled. Each prompt uses the model's chat template with a generation prompt appended and thinking disabled; the rendered text is encoded with special-token insertion disabled. Natural blocks and suffixes are sampled from the full softmax with temperature 1, top-$p = 1$, top-$k = -1$, min-$p = 0$, repetition penalty 1, and presence and frequency penalties 0. These settings override the model's generation-configuration defaults. Each request uses its own seed to generate one sequence. The answer-generation horizon is 8192 tokens, including the frozen block, so a continuation's budget is 8192 minus the block length. We replay blocks from their saved token ids; replica columns use separate generation requests. In the mathematical diagnostic, the candidate law is the renormalized top-20 token distribution, with ties broken by token id.

Table 5: Experimental panels and replication designs. Blocks are natural token sequences; suffix replicas use independent seeds.

Question	Panel	Replication design
Measurement noise	30 math problems, 60 prefixes	20 tokens $\times 16$ suffixes $\times 2$ pools
Natural capacity and feedback	24 ledger tasks, two models	16 blocks $\times 2$ ordinary suffixes
State–action propagation	Same 24 tasks, Qwen	16 blocks $\times 2$ readout suffixes
Selective reweighting	48 new ledger tasks, Qwen	8 blocks $\times(4+8)$ suffixes
SQL propagation	32 questions, two models	16 plans $\times 2$ suffixes
Independent SQL continuations	Same 32 questions, two models	16 plans $\times 8$ new suffixes
SQL replication	32 new questions, Qwen	16 new plans $\times(2+4)$ suffixes
SQL budget allocation	128 new questions, Qwen	16 plans $\times(2+8)$ suffixes

Table 6: Runtime limits and token ids passed to the sampler for stopping.

	Qwen	Gemma
Context window (tokens)	16,384	14,336
GPU memory fraction	0.88	0.94
Concurrent sequences	96	24
Stop token ids	248044, 248046	1, 50, 106, 248044, 248046

EOS is absorbing. The sampler uses the union of the engine's default stop ids 248044 and 248046, the model's generation-configuration EOS ids, and the tokenizer's EOS id; the stop-string list is empty. This gives two ids for Qwen and five for Gemma (Table 6). Neither inherited id, 248044 or 248046, occurs in the saved Gemma generations. All 1,024 original SQL continuations and 4,096 new continuations per model stop naturally: Qwen at 248046 and Gemma at 106 (Appendix I.1).

For a task x and m blocks, let Z_{br} contain its checkpoint outcomes in replica $r \in \{1, 2\}$. We use the symmetrized cross-column covariance

$$\widehat{G}_x = \frac{1}{2(m-1)} \sum_{b=1}^{m} \left[(Z_{b1} - \bar{Z}_1)(Z_{b2} - \bar{Z}_2)^\top + (Z_{b2} - \bar{Z}_2)(Z_{b1} - \bar{Z}_1)^\top \right]. \tag{53}$$

We average the matrices with equal weight for each task. We use the same observations for categorical one-hot outcomes, binary successes, and linear feature projections, and retain the signs of finite-sample estimates. For a binary final outcome, we estimate the matching variance budget by averaging the two replica success rates and subtracting the cross-replica product averaged over distinct blocks. The ratio of mean capacity to mean budget therefore describes within-task generation variation, with task difficulty removed before averaging.

The SQL measurements in Figure 3 (top) and Table 3 use the original two-replica pools. The mathematical diagnostic uses the budgets described in Appendix H.2; estimator identities appear in Appendix E.1. The independent SQL reference uses eight additional continuations per plan (Appendix I.4).

For the original panels, confidence intervals use 20,000 resamples of the original cluster unit. The independent SQL continuations use 10,000 draws with a seed fixed before generation (Appendix I.1); each additional analysis of saved pools reports its resampling scheme alongside its results. Ledger resampling retains the declared task strata; SQL resampling draws from its 19 database clusters, keeping all questions and replica columns together. Contrasts and decompositions use common resampling draws. For each ratio, we also report the number of draws with a nonpositive denominator. Readouts of matrix components, masking, and target response are secondary and use pointwise intervals. The prespecified primary endpoints are SQL draft–final correct covariance, ledger early-record–action covariance, ORC-minus-random joint correctness, and coverage-feedback joint delivery. The independent SQL forecast comparison is specified in Appendix I.

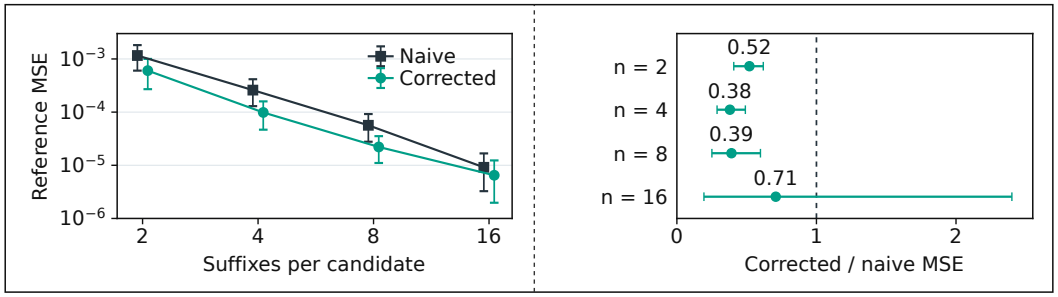

Figure 7: **Replication budget and measurement precision.** Estimates at 2, 4, 8 or 16 suffixes per candidate are compared with a shared independent reference pool on 60 math prefixes, each with 20 candidate tokens. Left: reference MSE of the uncorrected and corrected estimators. Right: corrected/naive MSE; the dashed line marks equal error. The 95% intervals resample the 30 problems, retaining both prefixes. The common reference-noise term cancels in the expected difference between estimator MSEs.

H.2 Independent measurement-noise diagnostic

The mathematical panel contains eleven AIME 2025, ten HMMT February 2025, and nine HMMT February 2026 problems from the MathArena releases. A generation is correct when it stops naturally and its final boxed answer passes verification against the dataset's reference answer; verification supports integer and noninteger reference answers. An answer truncated at the horizon scores zero and remains in the denominator. Each problem contributes its answer-root prefix and an interior prefix chosen at the highest candidate-law entropy in a middle window, preferring a line or sentence boundary. Each pool therefore contains 19,200 outcomes: 60 prefixes, 20 candidates, and 16 suffixes per candidate. The two pools use independent suffix seeds. Given frozen weights μ_a, the corrected estimator is

$$\widehat{\lambda}_{\mathrm{corr}} = \sum_a \mu_a \left(\widehat{q}_a - \sum_b \mu_b \widehat{q}_b \right)^2 - \sum_a \mu_a (1-\mu_a) \frac{s_a^2}{n_a}, \tag{54}$$

where s_a^2 has denominator $n_a - 1$. The second term estimates the entire suffix-noise inflation in the first term. At each budget $n = 2, 4, 8, 16$, we draw two hundred fixed-seed subsamples from the discovery pool and compare their estimates with the corrected estimate from the complete independent evaluation pool. These subsamples show how estimates from the existing pool vary with the budget. Both squared errors contain the same additive reference-estimation variance, which cancels in their expected difference.

Table 7: Error against the same independent corrected reference. The signed MSE difference is naive minus corrected, with a problem-cluster interval.

Suffixes/candidate	Corrected/naive MSE	MSE difference [95% interval]
2	0.520	5.572×10^{-4} $[2.968, 8.622] \times 10^{-4}$
4	0.380	1.609×10^{-4} $[0.756, 2.687] \times 10^{-4}$
8	0.391	3.443×10^{-5} $[1.317, 6.244] \times 10^{-5}$
16	0.709	2.671×10^{-6} $[-5.817, 11.551] \times 10^{-6}$

At $n = 16$, the mean naive variance across the panel is .002221, its estimated noise component is .001769, and the corrected result is .000452 $[-.000100, .001033]$. The noise-to-naive ratio is .797 $[.619, 1.061]$; its interval accounts for variation in both estimated components. The panel evaluates measurement precision when continuation noise accounts for most of the observed spread. On the ledger panel, the theoretical U-statistic and the implemented column covariance satisfy the finite-sample bridge identity on every task; their Qwen values, averaged with equal task weights, are .0355903 and .0361111, with difference $-.0005208$ $[-.0026042, .0011285]$.

H.3 SQL PLANNING AND EXECUTABLE OUTCOME CATEGORIES

The Spider development panel is selected before generation. Eligibility excludes repeated database–question–reference triples and requires official hard or extra-hard status, at least two distinct base tables in the parsed reference, and at least one aggregation, grouping, nested-query or set-operation construct. Both the shipped database and official distilled suite must be available. The reference query must pass the execution comparator against itself, and the rendered prompt must leave room for the complete 8192-token answer horizon within both models' context limits. These rules retain 214 of 1,034 development examples. Databases and their questions follow a fixed deterministic order; a round robin takes at most four questions per database until 32 are selected. The resulting panel contains 24 extra-hard and eight hard questions from 19 databases. Thirteen databases contribute two questions each and six contribute one. Four training examples, from separate databases, support engineering checks.

Each prompt requests `PLAN`, `DRAFT_SQL`, `CHECK`, and `FINAL_SQL` sections. We sample sixteen plans per question, stopping each at the first `END_PLAN` token boundary, EOS, premature SQL marker, or a 512-token block cap. Qwen contributes 504 marker-stopped, six cap-stopped, and two premature-SQL blocks; each stops before a complete SQL query is emitted. Every planned sampling slot remains in the denominator. Two continuations from each plan yield 1,024 paths per model, and two ordinary single-query roots per question give 64 comparator paths per model. Section markers must occupy their own lines. Draft extraction reads text up to the first `END_DRAFT`; final extraction reads text up to the first `END_FINAL`. A checkpoint is available when it yields a uniquely extractable, closed, single read-only statement; other checkpoints enter the unavailable category. Qwen supplies 1,022 extractable drafts and 1,020 finals; Gemma supplies 1,024 of each.

The five outcome coordinates are unavailable query, execution error, disagreement with the reference on the original database, original agreement followed by expanded-suite disagreement, and expanded-suite agreement. The official execution comparator runs each extracted query separately on the shipped database and on the distilled suite. It retains `DISTINCT` and uses the predicted literal values throughout execution. Queries execute against read-only database copies with a 60-second query–database limit. We stack the two checkpoint one-hot vectors into a ten-dimensional observation. We compute the full matrices and their linear projections.

Table 8: Additional SQL measurements. Intervals resample database clusters.

Quantity	Qwen	Gemma
Draft capacity λ_D	$0.0418\ [0.0187, 0.0657]$	$0.0141\ [0.0000, 0.0328]$
$\lambda_F - \lambda_D$	$-0.0034\ [-0.0194, 0.0134]$	$0.0043\ [-0.0052, 0.0150]$
Final correctness	0.5898	0.7324
Ordinary-root correctness	0.6406	0.7031
Workflow final minus root	$-0.0508\ [-0.1289, 0.0212]$	$0.0293\ [-0.0133, 0.0842]$
Draft-to-final correctness change	$0.0205\ [0.0019, 0.0444]$	$0.0137\ [-0.0334, 0.0852]$
Unique checkpoint queries	843	243

A companion input-structure analysis appears in Appendix K. It excludes complete questions and retains the original benchmark scores within each reported subset.

The final-capacity budget ratio for Qwen is $.46094\ [.26422, .64706]$. Gemma's raw final capacity is directly estimable, and its ratio point estimate is $.5402$; two bootstrap draws have zero budget, leaving its interval undefined under the declared rule. A within-task re-pairing reference permutes the entire second-replica vector. Exact enumeration by grouped integer counts gives upper-tail probabilities 6.98×10^{-10} and 1.01×10^{-5} for the draft–final statistic. These probabilities concern the frozen panel's shared-block independence; database-cluster intervals describe variation across the observed databases. Qwen has positive draft–final entries in 14 tasks across 11 databases; Gemma's three positive tasks come from two databases.

Among final queries that pass on the original database, 239/843 on Qwen and 172/922 on Gemma disagree on the expanded suite, giving pooled rates $.2835$ and $.1866$. The within-question conditional rates, averaged with equal question weights, are $.2902$ over 32 Qwen questions and $.1734$ over 31 Gemma questions with at least one original pass. These rates use different weighting conventions;

the main text reports the pooled counts. The selected `world_1_0766` illustration spans original-wrong and suite-wrong finals across 16 marker-stopped plans. Three block pairs are original-wrong and thirteen are suite-wrong, producing Equation (9). This post hoc example illustrates the matrix projection; panel-level category entries summarize all tasks. Executing saved queries on the official databases reveals a further distinction in this example: some generated queries filter countries by identifier whereas the reference filters by name; repeated country names in the expanded suite can make these differ. The matrix measures variation in reference-agreement categories under these execution tests. A separate car-manufacturer witness isolates a literal-value dependency: hard-coding Europe as continent identifier 2 agrees on the original database and fails after the expanded suite remaps identifiers. A network-query witness isolates repeated join contributions to an average; the prompt metadata abbreviates a composite primary key, while the actual SQLite schema and duplicate rows identify the aggregation unit.

H.4 Ledger generation and scoring

A ledger task gives a warehouse scope, an event table, pending transfers and reservations, open demand, pack sizes, and ten numbered rules. The executable reference applies those rules in order: select effective event revisions (R1–R4), compute item–location balances (R5), attempt transfers and subtract reservations (R6–R7), then determine shipment, unmet demand and replenishment (R8–R10). Each retained event within scope contributes its effective row; receipts and inbound adjustments add stock, while issues and outbound adjustments subtract it. A transfer executes in full only if the current source balance covers its quantity. Later transfers see the updated balances. Reservations are subtracted with a floor at zero. For each item, shipment is the minimum of allocable stock and open demand; replenishment is the smallest whole multiple of the pack size covering the remaining unmet demand.

The two void conventions differ at the revision-selection step. Both use the highest in-scope revision if it is posted. Under the eliminating convention, a void at that revision removes the event entirely. Under the fallback convention, it selects the highest posted in-scope revision, eliminating the event only when none exists. Each task states its convention explicitly.

The seeded generator produces 24-row ledgers with two in-scope items at two locations, one transfer and one reservation. Between a quarter and three eighths of rows are out of scope by warehouse, item or location. Similar event identifiers and separated revisions make row selection the principal source of ambiguity. For each task, the generator also designates an in-scope event to omit and an out-of-scope row to admit, then computes the effect of each perturbation on state and action. The 24-task panel is balanced by void convention. The disjoint 48-task panel crosses that convention with whether the designated omission changes the delivered action or only the state, giving four strata of twelve. Each task record contains the complete inputs, reference records, designated perturbations, and stratum assignment.

In Part 1, the model emits `STATE_1=` with a JSON record of effective revisions, eliminated events and ignored rows, followed by `END_STAGE_1`. This is the natural block boundary. In Part 2, the model performs the remaining computation; the answer's last nonempty line must contain `FINAL_JSON=` with `state` and `action` objects. The state contains allocable totals, unmet demand and a Boolean coverage flag; the action contains shipment and replenishment quantities. The parser rejects duplicate keys, non-finite literals and incorrect field types. An optional `checks` object in the state is ignored; other unexpected fields are invalid. State correctness requires exact agreement with the reference state, action correctness requires agreement with both action quantities, and joint correctness requires both. Invalid deliveries remain in the denominator. Target agreement in the reweighting study compares the delivered object with the frozen consensus object obtained from root generations, independently of the reference calculator.

H.5 Ledger state, action, and executable replay

The natural ledger panel has 24 tasks, equally divided between the two void conventions in Appendix H.4. Each model samples 16 selection blocks per task. Blocks end at a semantic marker, EOS, or a 4,096-token cap; 765 of the 768 blocks across models reach the marker. Two ordinary suffixes per block give 768 paths per model. Terminal categories are invalid delivery, both state and

action wrong, state wrong/action correct, state correct/action wrong, and both correct. Projecting these categories yields validity, state correctness, action correctness, and joint success.

The Qwen cross-feature covariance for validity, state correctness, and action correctness is

$$\begin{pmatrix} .00226 & .00408 & .00234 \\ .00408 & .03646 & .02856 \\ .00234 & .02856 & .02899 \end{pmatrix}.$$

The state–action entry is $.02856$ $[.01181, .04757]$. In the categorical matrix, the both-wrong/both-correct entry is $-.02448$ $[-.04358, -.00781]$. Feature projections reproduce the directly estimated feature matrix. For normalized category spectra, no task in either model satisfies the full five-category stability rule in the empirical-support diagnostic.

The temporal follow-up reuses the Qwen blocks and adds a `STATE_2` record of R5 cell balances before transfers and reservations. It adds 762 suffixes on 381 active blocks, while the three blocks that fail to reach the boundary retain their six outcome slots. A separate comparator uses 192 suffixes under the original contract, and a separate pilot uses 16 calls. The early-record and final-action diagonal capacities are $.04583$ $[.02066, .07188]$ and $.03681$ $[.01580, .05833]$; their paired difference is $-.00903$ $[-.03247, .01007]$. Early–late covariance is $.034028$ $[.013281, .055382]$. On the matched cohort, the readout contract changes joint delivery by $.0208$ $[-.0417, .0781]$ and adds 156 tokens $[-53, 361]$ on average.

Among paths with valid final delivery and an interpretable early record, 577 have correct early balances and 166 have wrong balances. Applying the deterministic calculator to those recorded balances reproduces all 743 final states and 742 final actions. Of the 166 wrong-balance paths, three still have correct final state and action, 70 have wrong state and correct action, and 93 have both wrong. The observed many-to-one shipment rule explains why these 70 state errors leave the action correct. Conditioned on the original block, these checkpoint relations show how the shipment rule maps early balances to final actions under the readout contract.

H.6 Selective reweighting and target alignment

The paired policy comparisons in Table 4 use the common budget and selection rules below.

Forty-eight fresh tasks form four structural strata of twelve. Eight complete root generations per task form pool C, from which we select and freeze a consensus state-and-action object. Eight independent natural blocks form pool B, with uniform law $\mu_b = 1/8$. Pool D provides four suffixes per block to estimate the target-conditioned means, rank tasks, and learn the exponential tilt. Each selector chooses three tasks per stratum. Pool E contains eight independent suffixes per block, or 3,072 evaluation paths. We evaluate each policy by weighting these outcomes with its frozen block weights. Random allocation is integrated exactly with task inclusion probability $1/4$. A separate 64-path direct-sampling check evaluates uniform and tilted laws on a fixed subset.

The consensus target is the modal canonical state–action object among valid pool-C answers that end before the horizon. Canonical JSON uses sorted keys and compact serialization; a fixed deterministic ordering of canonical objects breaks ties. Tasks with an empty valid-object pool retain uniform block weights. For each task, let k_b count target matches among block b's four discovery continuations, $\hat{q}_b = k_b/4$, and $\bar{q} = \frac{1}{8}\sum_b \hat{q}_b$. Its target-agreement capacity estimate is

$$\widehat{\lambda}_R = \frac{1}{8}\sum_b (\hat{q}_b - \bar{q})^2 - \frac{7}{256}\sum_b s_b^2, \qquad s_b^2 = \frac{4}{3}\hat{q}_b(1-\hat{q}_b).$$

All selectors share $g_b = \tilde{q}_b - \frac{1}{8}\sum_j \tilde{q}_j$, where $\tilde{q}_b = (k_b + \frac{1}{2})/5$, and weights $\pi_b \propto \exp(\beta g_b)$. Bisection chooses $\beta \geq 0$ for $D_{\mathrm{KL}}(\pi\|\mu) = 0.05$ using natural logarithms; constant directions remain uniform. Naive variance uses the first term of $\widehat{\lambda}_R$, outcome uncertainty uses $\bar{q}(1-\bar{q})$, and predicted response uses $\sum_b(\pi_b - \frac{1}{8})\tilde{q}_b$. Capacity summaries retain signed estimates; ORC and naive-variance ranking scores are floored at zero. Each selector takes its top three tasks per stratum, with score ties broken by a fixed deterministic ordering of task identifiers.

Let $d_x^Y = \sum_b (\pi_x(b) - \mu_x(b))\widehat{m}_{x,Y}(b)$ and let I_x indicate ORC selection. The all-task ORC-minus-random estimate is

$$\widehat{\Delta}_Y = \frac{1}{48}\sum_x (I_x - 1/4)d_x^Y. \tag{55}$$

To calculate a group's additive contribution, we restrict the sum to that group and retain the denominator 48. Hence the correct-target contribution .0058297 and wrong-target contribution $-.0031273$ sum to the all-task correctness contrast .0027024. The corresponding wrong-target contribution to agreement is $+.0044053$. After policy selection, we determine group membership by comparing the frozen target with the oracle; all five wrong-target tasks are retained. Four are selected by ORC, and four contain a wrong state with the correct action.

Table 9: Frozen policy values on the common evaluation pool. R denotes agreement with the consensus target, V valid delivery, S correct state, L correct action, and J joint correctness.

Selector	R	V	S	L	J
ORC	0.8162	0.9842	0.7855	0.9069	0.7852
Block surprisal	0.8064	0.9831	0.7826	0.9016	0.7823
Outcome uncertainty	0.8134	0.9832	0.7827	0.9057	0.7824
Naive variance	0.8151	0.9831	0.7844	0.9059	0.7841
Predicted response	0.8151	0.9831	0.7844	0.9059	0.7841
Random allocation	0.8060	0.9839	0.7828	0.9028	0.7825
Uniform blocks	0.7979	0.9837	0.7777	0.8997	0.7773
Tilt all tasks	0.8303	0.9843	0.7982	0.9119	0.7978

The predeclared block-entropy comparator uses mean sampled sequence surprisal, summed over retained tokens. Its correlation with block length is .898 and with ORC score .243. Thus this comparator measures a sequence-level quantity. ORC and naive variance share eleven of their twelve selected tasks. The target endpoint R is secondary. The equal-total-KL diffuse allocation is a post hoc comparison constructed after inspecting pool E. Nonconstant directions achieve KL .05; six task directions are constant. Realized summed KL is .600 for ORC, .525 in expectation for random allocation, .500 for block surprisal, and 2.100 for tilting every nonconstant task. The original selectors choose the same number of tasks and use the same observations and discovery effort. This diffuse allocation spreads total KL .600 over the 42 nonconstant directions and gives joint gain .01141 $[.00564, .01717]$ over uniform blocks; its direct comparison with ORC is $[-.00218, .00922]$.

Discovery and evaluation target capacities have Pearson correlation .818 and Spearman correlation .636. The mean held-out capacity is .0874 on ORC-selected tasks and .0157 on the remaining tasks. Discovery capacity correlates .796 with held-out target response and .298 with correctness response. Thirty-seven of 42 learned nonconstant tilts improve held-out target agreement. The response experiment holds the block library and suffix policy fixed and forms the consensus target from model-generated answers.

H.7 Coverage feedback as an expanded policy

The deterministic gate reads source rows and the model's emitted selection record. It reports omitted in-scope events, impossible revisions, invented identifiers, and positive selections lacking in-scope source support. It leaves the choice among existing revisions to the model. Of 384 blocks per model, the gate flags 220 on Qwen and 126 on Gemma. A prespecified hash ordering deterministically selects 128 flagged Qwen blocks. These and all 126 flagged Gemma blocks receive two suffixes in each review arm: generic review or the same wrapper containing the specific findings. Both are introduced in a new user turn. Unflagged and unobservable blocks follow ordinary continuation.

For Qwen, we report a cohort-standardized contrast: the mean repair-minus-ordinary difference over the 128 selected blocks, multiplied by the full panel's flagged fraction $220/384$. Equivalently, each selected difference receives expansion factor $220/128$ before averaging over all sixteen block slots and all tasks. This factor defines the standardization. Interpreting the result as the effect of feedback on all 220 flagged blocks assumes equal mean feedback effects in the selected and remaining flagged blocks. Gemma includes all flagged blocks and uses factor one. The task bootstrap and nested-prefix sensitivity analysis resample observed tasks and, in the nested analysis, prefixes within the selected

cohort. Endpoints and contrasts share resampling draws, with expansion factors fixed. Under ordinary continuation, joint success is 598/768 on Qwen and 724/768 on Gemma. We obtain the reported standardized levels by adding these contrasts to the ordinary-continuation baselines.

Table 10: Joint-delivery contrasts relative to ordinary continuation. Qwen uses cohort standardization; Gemma includes all flagged blocks. All numbers are percentage points.

Model	Feedback	Task-cluster estimate [95% CI]	Nested-prefix CI
Qwen	Generic	$2.014\ [-1.567, 5.371]$	$[-2.686, 6.714]$
Qwen	Coverage	$5.595\ [1.567, 9.176]$	$[0.671, 10.518]$
Qwen	Coverage minus generic	$3.581\ [0.895, 6.043]$	$[-0.224, 7.610]$
Gemma	Coverage	$-0.651\ [-1.823, 0.391]$	—

Table 11: Additive contributions to the coverage-feedback contrast, classified by replaying each frozen record before intervention. Qwen uses cohort standardization; the denominator remains the complete panel. Contributions are percentage points.

Replay group	Qwen blocks	Gain	Gemma blocks	Gain
Exact record	121	0	134	0
Different record, same computed state and action	209	2.462	247	−1.042
Computed state differs, action same	13	0.448	1	0
Computed action differs	34	2.686	2	0.391
Unreadable record	7	0	0	0
Total	384	5.595	384	−0.651

The replay-unchanged group contributes $.02462\ [-.01119, .05819]$ on Qwen and $-.01042\ [-.02083, -.00130]$ on Gemma. Its Qwen contribution is 44% of the total point estimate. Inference uses absolute contributions to keep a common scale across bootstrap draws, including those whose total crosses zero. The action-changing group's Qwen contribution is $.02686\ [.00671, .04924]$, with nested interval $[.00224, .05819]$.

We read only public records newly emitted after the frozen boundary. On the matched 128-block Qwen cohort, ordinary, generic, and coverage arms emit 58, 238, and 219 valid new records out of 256 paths each. Under the latest-visible-record rule, which inherits the frozen record when no new attempt appears, the cohort-standardized coverage contrast in exact-record correctness is $.1365\ [.0918, .1835]$ and the contrast in the record content used by the task is $.0380\ [.0112, .0694]$. Their paired difference is $.0985\ [.0515, .1499]$. These record readouts show how the evidence supplied to the final computation changes, complementing the complete-delivery outcome. The feedback experiment changes the suffix policy through a new user turn. Its observed-cohort contrasts and standardized contributions complement the fixed-kernel block-reweighting response experiment.

H.8 Measurement budget and reuse

The main cost is generating independent continuations. With m sampled blocks and n suffixes per block, the design makes mn suffix calls; encoding another checkpoint already observed in those continuations reuses the same calls. SQL extracts draft and final results from the same 2,048 continuations across the two models. Its five categories at each checkpoint are computed by query execution. The ledger temporal study adds an explicit readout to the continuation contract and accounts separately for its 762 new continuations on active blocks.

For each task in the reweighting experiment, eight root answers form the target, eight generated blocks define the choice library, and 8×4 discovery plus 8×8 evaluation continuations estimate the responses. The same 3,072 evaluation paths score all frozen selectors. Comparing an additional selector on this pool requires only weighted arithmetic. Repeated applications can share the policy's discovery cost; a new task requires its own discovery samples.

Figure 7 quantifies a direct budget–precision tradeoff: at four suffixes per candidate the corrected/naive reference-MSE ratio is 0.380 using identical generation data. The original SQL cost record lists

1.713 rented H100 hours and USD 15.16 including model transfer; the reweighting experiment used 8.672 rented hours and USD 35.90. The independent SQL pool adds 8,192 continuations on the same frozen plans at a recorded cost of USD 10.94. Recorded costs include experimental overhead; call counts and estimator errors quantify the cost and precision of these reusable measurements. The additional ledger and semantic-sensitivity analyses reuse saved model outputs.

H.9 Separating plan and continuation covariance

A draft query and its final version share both the plan and the subsequent generation. Let D and F denote their binary execution outcomes and B the sampled plan, with the task and initial prefix fixed. The covariance along a single continuation decomposes as

$$\mathrm{Cov}(D, F) = \underbrace{\mathrm{Cov}_B(\mathbb{E}[D \mid B], \mathbb{E}[F \mid B])}_{\Gamma_{DF}} + \mathbb{E}_B[\mathrm{Cov}(D, F \mid B)]. \tag{56}$$

The first term captures how plans jointly shift the two outcome laws. The second captures the relation between draft and final outcomes within continuations of a fixed plan. Independent replicas separate these sources: if (D_{b1}, F_{b1}) and (D_{b2}, F_{b2}) are the two continuations of plan b, define

$$\widehat{C}_{\mathrm{same}} = \tfrac{1}{2}\{\widehat{\mathrm{Cov}}(D_1, F_1) + \widehat{\mathrm{Cov}}(D_2, F_2)\}, \tag{57}$$

$$\widehat{\Gamma}_{DF} = \tfrac{1}{2}\{\widehat{\mathrm{Cov}}(D_1, F_2) + \widehat{\mathrm{Cov}}(D_2, F_1)\}. \tag{58}$$

Each sample covariance is computed across the 16 plans, with denominator 15. Their difference satisfies the exact finite-sample identity

$$\widehat{C}_{\mathrm{same}} - \widehat{\Gamma}_{DF} = \tfrac{1}{2}\widehat{\mathrm{Cov}}(D_1 - D_2, F_1 - F_2), \tag{59}$$

and estimates the second term in Equation (56).

Table 12: Draft–final covariance in SQL. Values are means with equal weight for each question; 95% intervals resample the 19 databases, using the same draws for all three quantities.

Quantity	Qwen	Gemma
Same continuation	$0.06960\ [0.04397, 0.09482]$	$0.01484\ [0.00119, 0.03321]$
Shared plan	$0.03411\ [0.01597, 0.05363]$	$0.01152\ [-0.00035, 0.02874]$
Conditional continuation	$0.03548\ [0.01746, 0.05576]$	$0.00332\ [0.00000, 0.00753]$

For Qwen, the within-plan covariance is comparable in size to the between-plan term. The observed relation across checkpoints therefore has two substantial sources: shifts in future outcome probabilities across plans and dependence between draft and final outcomes within a fixed plan.

Branch spread with two continuations. The same replica design measures how sampling continuations contributes to the estimated variation across branches. For either checkpoint outcome Y, write $\bar{Y}_b = (Y_{b1} + Y_{b2})/2$. The naive across-plan variance and the cross-replica estimate satisfy

$$\widehat{\mathrm{Var}}_b(\bar{Y}_b) - \widehat{\mathrm{Cov}}_b(Y_{b1}, Y_{b2}) = \tfrac{1}{4}\widehat{\mathrm{Var}}_b(Y_{b1} - Y_{b2}). \tag{60}$$

The expected right-hand side is $\frac{1}{2}\mathbb{E}_B[\mathrm{Var}(Y \mid B)]$, the contribution of continuation sampling to the variance of the two-replica branch mean.

The excess varies across questions, so correction changes both the estimated magnitude and the ranking of branch capacities. These comparisons quantify the contribution of continuation noise in the SQL panel. The mathematical-reasoning study in Appendix H.2 measures estimation error against an independent reference with many continuations. All summaries retain the signs of covariance estimates, and tied task estimates receive average ranks. The tables are computed directly from the recorded checkpoint scores.

H.10 Matched-budget selection and target alignment

Selecting among learned nonconstant directions. A random selector with a matched budget isolates the effect of ranking after constant learned directions have been screened out. The four strata

Table 13: Across-plan spread at two continuations per plan. "Excess" is the paired difference in Equation (60); intervals resample databases. Rank agreement is Spearman correlation between the 32 question-level naive and corrected estimates.

Model	Outcome	Naive	Corrected	Excess [95% interval]	Rank corr.
Qwen	Draft	0.07041	0.04180	0.02861 [0.01768, 0.04043]	0.753
Qwen	Final	0.06172	0.03841	0.02331 [0.01240, 0.03517]	0.710
Gemma	Draft	0.01986	0.01406	0.00579 [0.00000, 0.01398]	0.908
Gemma	Final	0.02572	0.01836	0.00736 [0.00215, 0.01473]	0.647

Table 14: Incorrect-target contributions to the full 48-task ORC-minus-random comparison, in percentage points.

Task	ORC selected	Contribution to R	Contribution to J
eliminate_05	Yes	$+0.116$	-0.153
eliminate_08	No	-0.020	$+0.017$
eliminate_18	Yes	$+0.149$	-0.028
eliminate_20	Yes	$+0.093$	-0.077
fallback_14	Yes	$+0.102$	-0.073
Total	4 of 5	$+0.441$	-0.313

contain 10, 10, 11, and 11 nonconstant directions. The selector chooses three of these directions uniformly in each stratum and applies the original KL-0.05 tilt to each selected task. Every selection therefore uses total KL 0.600, matching ORC. Exact averaging over this selection law gives inclusion probability $3/n_s$ for each eligible task in stratum s; all 48 tasks contribute to the reported policy values. A constant learned direction leaves its block law unchanged, independently of the task's population capacity.

ORC exceeds this matched-budget random selector by 0.908 $[0.190, 1.694]$ percentage points in target agreement and 0.208 $[-0.504, 0.955]$ points in joint correctness. Their paired difference is 0.700 $[0.077, 1.488]$ points. Thus the target-agreement advantage remains after matching both the number of selected tasks and realized KL. As in the original comparisons, intervals use the same 20,000 stratified task draws, keeping each task's frozen target, block library, direction, and policy values together.

The five incorrect targets. Table 14 breaks down the original ORC-minus-random contrast into individual task contributions. Each entry is the task's policy difference divided by 48 and expressed in percentage points. On each of the four selected incorrect-target tasks, reweighting increases agreement with the frozen target while reducing joint correctness. The unselected task contributes in the opposite direction because the random comparator sometimes selects it and tilts toward its target. These five contributions sum to $+0.441$ points in agreement and -0.313 points in joint correctness.

To measure each task's influence, we omit each incorrect-target task in turn while retaining the four strata's equal weights and all remaining task-level policies. The full-panel agreement contrast ranges from 0.890 to 1.074 points and the correctness contrast from 0.253 to 0.435 points. The remaining incorrect-target contribution to correctness is negative in all five leave-one-task-out calculations; the main estimates retain all 48 tasks.

I Independent continuations and held-out response

I.1 A frozen SQL intervention evaluated on fresh continuations

We extend the SQL panel of Appendix H.3 by replaying its original plans and sampling fresh continuations. The panel contains 32 questions from 19 databases, with 16 plans per question and model. Each plan receives eight new continuations, giving 4,096 per model and 8,192 in total. We use the original two continuations to determine the intervention and its two forecasts before observing

any fresh outcome. Qwen is the primary model; Gemma is a prespecified replication on the same question panel. We keep the models separate throughout the analysis.

The new continuations start from the stored plan token ids, with disjoint random seeds and the same models, sampling settings, stop sets, and 8,192-token answer horizon. All fresh continuations stop before the horizon and are included in the evaluation. The original and fresh pools use the same expanded-suite correctness criterion. The plan library, original scores, weights, forecasts, endpoint, and resampling settings were fixed before fresh generation. A subsequent CPU-only audit covered all 26,880 historical checkpoints and re-executed all available saved SQL. The audit corrected 98 execution-probe timeouts: 45 correctness labels changed from zero to one, and 53 changed only their error category. All other labels were reproduced. The primary analysis retains the original frozen weights and evaluates forecasts and responses with the corrected final labels.

For one question and model, let $D_{ir}^{\mathrm{freeze}} \in \{0,1\}$ denote the draft label recorded when the weights were frozen, and let $F_{ir} \in \{0,1\}$ denote the corrected final correctness of plan i in original replica $r \in \{1,2\}$. Each original draft column defines a direction: every plan whose draft succeeded receives twice the weight of a plan whose draft failed. With $m = 16$, uniform reference law $\mu_i = 1/m$, and $\bar{D}_r^{\mathrm{freeze}} = m^{-1} \sum_i D_{ir}^{\mathrm{freeze}}$, the weights and contrast are

$$\begin{aligned} \pi_{ir} &= \frac{2^{D_{ir}^{\mathrm{freeze}}}}{\sum_j 2^{D_{jr}^{\mathrm{freeze}}}}, \qquad a_{ir} = \pi_{ir} - \frac{1}{m}, \\ a_{ir} &= \frac{D_{ir}^{\mathrm{freeze}} - \bar{D}_r^{\mathrm{freeze}}}{m(1 + \bar{D}_r^{\mathrm{freeze}})}, \qquad \sum_i a_{ir} = 0. \end{aligned} \tag{61}$$

This gives two directions per question. When a draft column is constant, $a_{\cdot r} = 0$. We retain all 64 directions per model, including 23 constant directions on Qwen and 55 on Gemma. Both directions are constant on 10 Qwen questions and 27 Gemma questions.

Let $\bar{F}_i^E$ be the mean final correctness over the eight fresh continuations of plan i, and let $r' \neq r$ be the other original replica. The same-continuation forecast, independent forecast, and observed response are respectively

$$s_r = \sum_i a_{ir} F_{ir}, \qquad c_r = \sum_i a_{ir} F_{ir'}, \qquad e_r = \sum_i a_{ir} \bar{F}_i^E. \tag{62}$$

Each compares the frozen intervention with uniform weights on the same plans. For the primary endpoint, we compare squared errors within each direction, then average the two directions within each question and give all 32 questions equal weight:

$$L = \frac{1}{32} \sum_{q=1}^{32} \frac{1}{2} \sum_{r=1}^{2} \left[(s_{qr} - e_{qr})^2 - (c_{qr} - e_{qr})^2 \right]. \tag{63}$$

Thus positive L favours the independent forecast. Intervals use 10,000 percentile bootstrap draws of the 19 database clusters, carrying all questions and directions in each selected cluster and recomputing the equal-question mean. The resampling seed was fixed before generation.

On Qwen, the intervention improves final correctness on independent continuations by 2.439 pp, while the same-continuation forecast predicts 4.485 pp. Using the other original replica gives 2.204 pp and reduces squared forecast error by 8.574 pp^2 $[3.149, 15.527]$ (Table 15). On Gemma, the independent gain is 0.806 pp and the loss difference is 0.128 pp^2 $[-0.197, 0.530]$. Most Gemma directions are constant, and the two forecasts are close.

In a separate sensitivity analysis of the finite suffix pool, we resample paired draft–final rows within each plan while holding the question panel fixed. This places L in $[6.584, 10.608]$ pp^2 on Qwen and $[-0.192, 0.432]$ pp^2 on Gemma. A question's two directions share the same resampled fresh plan means. These ranges describe variation from the finite fresh pool, conditional on the observed questions; the database-cluster intervals in Table 15 remain the intervals for the primary endpoint.

For a replay sensitivity check, we rebuild the original twice-the-weight rule from the corrected draft labels. This changes the weights on six Qwen questions and gives $L = 8.613$ $[3.189, 15.540]$ pp^2. The primary result above preserves the intervention that was fixed before fresh generation.

Table 15: Frozen SQL forecasts and independent outcomes. Each mean averages the two directions within each question and gives all questions equal weight. Brackets show 95% intervals from 10,000 database-cluster bootstrap draws. Qwen is the primary model; Gemma is the prespecified replication.

Quantity	Qwen	Gemma
Same-continuation forecast s (pp)	$4.485\ [2.795, 6.131]$	$0.911\ [0.063, 2.083]$
Independent forecast c (pp)	$2.204\ [1.077, 3.413]$	$0.719\ [-0.020, 1.878]$
Fresh outcome e (pp)	$2.439\ [1.412, 3.514]$	$0.806\ [0.030, 1.972]$
Signed error $s - e$ (pp)	$2.046\ [1.107, 3.044]$	$0.105\ [-0.012, 0.271]$
Signed error $c - e$ (pp)	$-0.235\ [-0.900, 0.372]$	$-0.087\ [-0.212, -0.0001]$
Mean $(s - e)^2$ (pp^2)	$13.374\ [6.338, 21.461]$	$0.348\ [0.038, 0.783]$
Mean $(c - e)^2$ (pp^2)	$4.800\ [1.870, 8.471]$	$0.220\ [0.003, 0.510]$
Paired difference L (pp^2)	$8.574\ [3.149, 15.527]$	$0.128\ [-0.197, 0.530]$

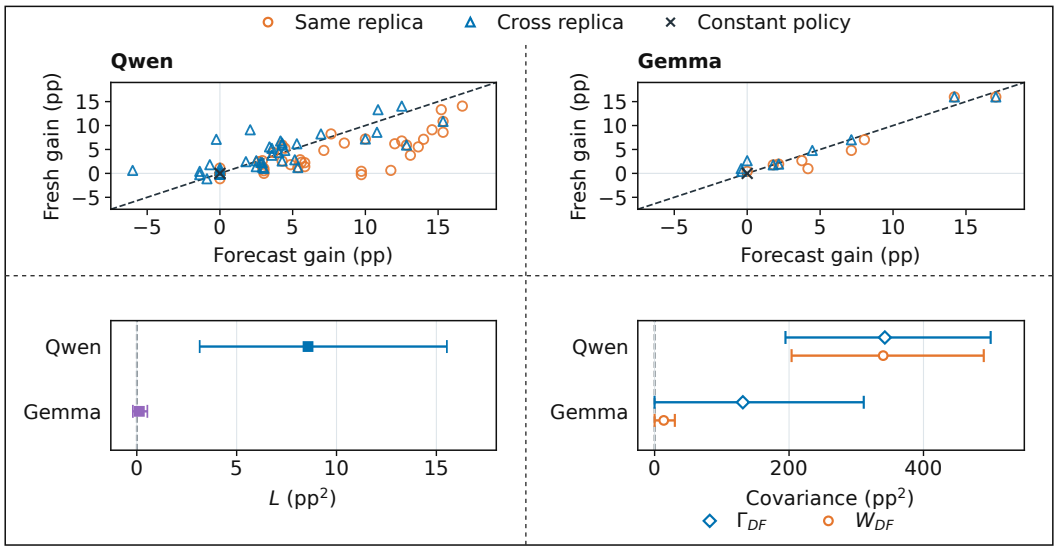

Figure 8: **Original-panel forecasts and covariance sources.** Top: forecast versus gain on fresh continuations for Qwen (left) and Gemma (right), showing all 64 directions per model; constant directions overlap at zero (Qwen: 23, Gemma: 55). Bottom left: paired loss difference L; positive values favor independent prediction. Bottom right: fresh-pool corrected shared-plan covariance $\widehat{\Gamma}_{DF}$ and within-continuation covariance W_{DF}. Bars show 95% database-cluster intervals.

I.2 Forecast sensitivity to input-defined SQL conventions

We apply the frozen forecast comparison to the two combined structural panels in Appendix K. Their question lists were defined from task inputs before the fresh-continuation experiment. The 15-question panel excludes questions involving aggregate extrema with a single winner and unspecified tie-breaking, country exclusions based on names, and relevant inconsistencies in key descriptions; the 10-question panel also removes the broader reference conventions listed there. For both models, we use the original retained question lists, unchanged weights, corrected execution labels, and all directions, including constant ones. These input-defined subsets provide sensitivity analyses alongside the primary 32-question panel.

On Qwen, the independent forecast continues to have lower squared error on both subsets: $L = 5.001\ [1.079, 11.708]\ \text{pp}^2$ on the 15-question panel and $1.764\ [0.381, 3.642]\ \text{pp}^2$ on the 10-question panel. These subsets establish the forecast advantage across the retained reference conventions. Removing any one database from the full Qwen panel leaves a positive point estimate, with L ranging from 6.081 to $9.162\ \text{pp}^2$. Together, these within-panel analyses quantify sensitivity to task conventions and database composition.

Table 16: Frozen SQL forecasts on the original structural subsets. T/DB denotes questions/databases, and NC counts nonconstant directions out of two per question. Gains and forecasts are in pp; L is in pp^2. Brackets give pointwise 95% intervals from 10,000 database-cluster draws. Each draw averages the two directions within a question and retains equal question weighting.

Panel (T/DB)	Model	NC	e	s	c	L
Full (32/19)	Qwen	41/64	2.439 [1.412, 3.514]	4.485 [2.795, 6.131]	2.204 [1.077, 3.413]	8.574 [3.149, 15.527]
Full (32/19)	Gemma	9/64	0.806 [0.030, 1.972]	0.911 [0.063, 2.083]	0.719 [−0.020, 1.878]	0.128 [−0.197, 0.530]
Combined (15/11)	Qwen	19/30	1.321 [0.217, 2.693]	3.025 [0.889, 5.413]	1.428 [0.461, 2.599]	5.001 [1.079, 11.708]
Combined (15/11)	Gemma	7/30	0.654 [0.000, 1.616]	0.901 [0.000, 2.051]	0.492 [−0.057, 1.527]	0.273 [−0.420, 1.115]
Extended (10/8)	Qwen	11/20	1.109 [0.070, 3.070]	2.216 [0.349, 5.535]	0.936 [0.208, 2.092]	1.764 [0.381, 3.642]
Extended (10/8)	Gemma	0/20	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]

Gemma's 15-question loss contrast is 0.273 $[-0.420, 1.115]$ pp^2. All 20 directions in its 10-question subset have $a_{ir} = 0$, giving zero forecasts, response, and loss difference by construction.

The analysis uses all 2,048 original and 8,192 fresh continuation labels, with the same fixed weights and question subsets.

I.3 What the forecast reads

The finite-library response in (5) is $\sum_i a_{ir} m_F(i)$, where $m_F(i) = \mathbb{E}_K[F \mid i]$. For the empirical covariance Cov_μ over the uniform plan library, substitution of (61) gives the exact identities

$$s_r - e_r = \frac{\mathrm{Cov}_\mu(D^{\mathrm{freeze}}_{\cdot r}, F_{\cdot r} - \bar{F}^E_{\cdot})}{1 + \bar{D}^{\mathrm{freeze}}_r}, \tag{64}$$

$$c_r - e_r = \frac{\mathrm{Cov}_\mu(D^{\mathrm{freeze}}_{\cdot r}, F_{\cdot r'} - \bar{F}^E_{\cdot})}{1 + \bar{D}^{\mathrm{freeze}}_r},$$

$$(s_r - e_r)^2 - (c_r - e_r)^2 = (s_r - c_r)(s_r + c_r - 2e_r). \tag{65}$$

The forecast error therefore depends on how the draft-defined direction aligns with the final-outcome residual. The loss difference also depends on how both forecasts compare with the independent response.

Assume that continuation–recorded-label pairs are independent given the fixed plans and that final outcomes follow a common per-plan law. Expectations below include any randomness in recording the draft labels. Conditioning on the library and on the single original draft column used to form the weights gives

$$\mathbb{E}[c_r \mid \text{library}, D^{\mathrm{freeze}}_{\cdot r}] = \mathbb{E}[e_r \mid \text{library}, D^{\mathrm{freeze}}_{\cdot r}] = \sum_i a_{ir} m_F(i). \tag{66}$$

Conditioning on the single weight-defining draft column fixes the weights while preserving the randomness of the other replica and the fresh pool. For the same-continuation forecast, let $m^{\mathrm{freeze}}_D(i) = \mathbb{E}[D^{\mathrm{freeze}} \mid i]$ and $\beta_i = \mathrm{Cov}(D^{\mathrm{freeze}}, F \mid i) / \mathrm{Var}(D^{\mathrm{freeze}} \mid i)$ where the denominator is positive. Binary regression gives

$$\mathbb{E}[s_r \mid \text{library}, D^{\mathrm{freeze}}_{\cdot r}] - \sum_i a_{ir} m_F(i) = \sum_i a_{ir} \{D^{\mathrm{freeze}}_{ir} - m^{\mathrm{freeze}}_D(i)\} \beta_i, \tag{67}$$

with zero contribution for deterministic draft outcomes. This term describes how within-continuation dependence enters a forecast whose weights use that same continuation. The empirical loss difference in (63) measures squared-error performance alongside this conditional-mean identity. A common additive change in all plans' final success probabilities cancels because $\sum_i a_{ir} = 0$. A plan-dependent change δ_i changes the response by $\sum_i a_{ir} \delta_i$.

I.4 Capacity and covariance on the fixed plan library

All capacity and covariance estimates in this subsection use the corrected execution labels for both draft and final outcomes. For a scalar outcome X, let $\bar{X}_i$ and s_i^2 be its plan mean and unbiased within-plan variance from n replicas. The fixed-library uncorrected and corrected capacities are

$$N_n = \frac{1}{m}\sum_i (\bar{X}_i - \bar{X})^2, \qquad C_{U,n} = N_n - \frac{m-1}{m^2 n}\sum_i s_i^2. \tag{68}$$

Conditional on the plans, $C_{U,n}$ is unbiased for $\mathrm{Var}_\mu \mathbb{E}_K[X \mid i]$. We compare the original $n = 2$ estimates with a fresh $n = 8$ reference, retaining signed estimates.

Table 17: Capacities for final correctness on the fixed 16-plan library, in pp^2. Both corrected rows use the within-plan noise correction in (68). Intervals use the same database-cluster draws as Table 15.

Estimator	Qwen	Gemma
Original uncorrected N_2	578.6 [369.0, 782.9]	241.1 [68.7, 452.0]
Original corrected $C_{U,2}$	368.0 [185.9, 562.9]	167.8 [3.4, 369.3]
Fresh corrected $C_{U,8}$	369.4 [208.7, 536.9]	170.1 [25.1, 353.5]

For final correctness, the ratio of the original uncorrected mean to the fresh corrected mean is 1.566 on Qwen and 1.417 on Gemma. The correction brings the average capacity close to the independent reference. The paired per-question squared-error difference, uncorrected minus corrected, is $-1{,}545$ pp^4 $[-156{,}112, 126{,}039]$ on Qwen and $6{,}777$ pp^4 $[-22{,}613, 39{,}124]$ on Gemma. Using the same independent reference for both estimates cancels its sampling-noise contribution in the expected paired difference.

The original cross-column estimate in (53) and the noise-corrected estimate in (68) are distinct finite-sample estimators. Multiplying an estimate with denominator $m-1$ by $(m-1)/m = 15/16$ converts it to the fixed-library scale. For a scalar outcome with two replicas, set $d_i = X_{i1} - X_{i2}$. Their exact finite-sample relation at this scale is

$$C_{\mathrm{col},2} - C_{U,2} = \frac{(\sum_i d_i)^2 - \sum_i d_i^2}{4m^2}. \tag{69}$$

For final correctness, the rescaled cross-column means are 360.107 and 172.119 pp^2, compared with 368.042 and 167.847 pp^2 for $C_{U,2}$ on Qwen and Gemma. The corresponding uncorrected-to-corrected ratios are 1.607 and 1.401 for the cross-column estimate, and 1.572 and 1.436 for $C_{U,2}$.

Draft–final dependence. Let $B_{DF} = \mathrm{Cov}_\mu(\bar{D}_i, \bar{F}_i)$ be the empirical covariance of fresh plan means, W_{DF} the average unbiased within-plan draft–final covariance, and T_{DF} the covariance of the mn pooled observations, with denominator mn. For each question,

$$T_{DF} = B_{DF} + \frac{n-1}{n} W_{DF}, \qquad \widehat{\Gamma}_{DF} = B_{DF} - \frac{m-1}{mn} W_{DF}. \tag{70}$$

The first identity holds for the observed sample. In the second expression, subtracting the correction term removes finite-replica covariance from the estimated plan means, yielding an unbiased estimate of the shared-plan draft–final covariance under conditional independence. Here $m = 16$ and $n = 8$, so the two coefficients are $7/8$ and $15/128$.

On Qwen, the within-continuation component $(7/8)W_{DF}$ accounts for 43.8% of the empirical pooled covariance. The corrected shared-plan covariance is 342.5 pp^2, while the within-continuation covariance is 340.1 pp^2. Gemma's within-continuation covariance is 13.6 pp^2 $[0.0, 30.1]$. This pattern is consistent with the larger gap between Qwen's forecasts. Equation (67) gives the direction-weighted link between continuation dependence and forecast error.

I.5 Continuation budget on the frozen SQL pool

The eight fresh continuations per plan allow us to examine budget effects within a fixed pool. For each question and each $n \in \{2, 4, 8\}$, independently choose a uniform n-element subset of the eight

Table 18: Draft–final covariance in the fresh SQL pool, in pp^2, with 95% database-cluster intervals. The empirical identity is $T_{DF} = B_{DF} + (7/8)W_{DF}$; the corrected plan covariance subtracts $(15/128)W_{DF}$ from B_{DF}.

Quantity	Qwen	Gemma
Covariance of plan means B_{DF}	382.4 [224.8, 545.6]	132.9 [0.1, 313.6]
Within-continuation covariance W_{DF}	340.1 [203.8, 489.7]	13.6 [0.0, 30.1]
Pooled covariance T_{DF}	679.9 [427.8, 927.9]	144.8 [4.0, 330.0]
Finite-replica term $\frac{15}{128}W_{DF}$	39.8 [23.9, 57.4]	1.6 [0.0, 3.5]
Corrected plan covariance $\widehat{\Gamma}_{DF}$	342.5 [194.7, 499.9]	131.3 [−0.1, 311.1]

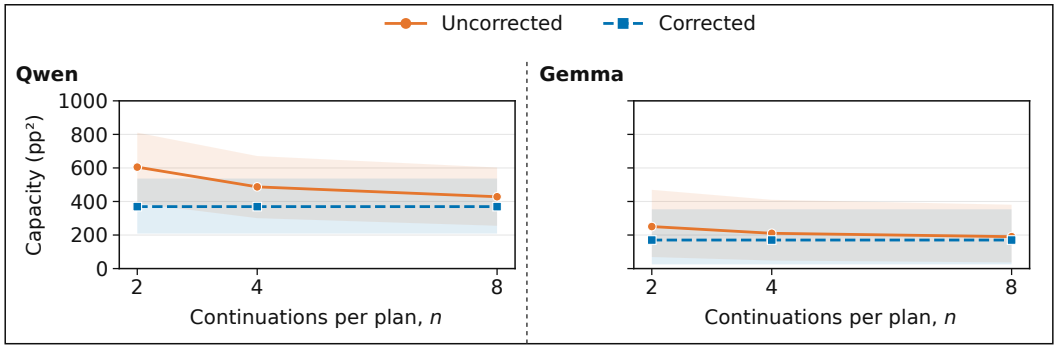

Figure 9: **Finite-pool continuation budget.** Uncorrected and signed corrected capacities, averaged exactly over subsets of the eight fresh continuations per frozen plan. Both panels use the same axes; shading denotes 95% database-cluster intervals conditional on the observed pool. The corrected line follows the exact subset-average identity in Equation (72); the vertical separation shows the contribution of suffix noise removed by the correction.

continuations within every plan, and average Equation (68) over all such choices. This keeps the 16-plan library and the scoring rule fixed. We compute the subset averages exactly from moments and check them by enumerating all subsets within each plan. Inference retains the question and database sampling units.

Let N_8 be the variance of plan means computed from the full pool, with denominator $m = 16$, let $s^2_{i,8}$ be the unbiased variance within plan i, and define $A = (m-1)m^{-2}\sum_i s^2_{i,8}$. Writing $\mathbb{E}_{\mathrm{sub}}$ for averaging over subsets conditional on the observed pool, sampling without replacement gives

$$\mathbb{E}_{\mathrm{sub}}[N_n] = N_8 + \left(\frac{1}{n} - \frac{1}{8}\right) A = C_{U,8} + \frac{A}{n}, \tag{71}$$

$$\mathbb{E}_{\mathrm{sub}}[C_{U,n}] = C_{U,8}. \tag{72}$$

A subset mean in plan i has variance $(1/n - 1/8)s^2_{i,8}$, and its sample variance has expectation $s^2_{i,8}$. Independence of the subset choices across plans then yields these identities.

Qwen's uncorrected-to-corrected ratio falls from 1.638 at $n = 2$ to 1.159 at $n = 8$; Gemma's falls from 1.473 to 1.118. The measured A determines this budget dependence, and the correction removes it exactly in the subset average. Intervals use 10,000 resamples of the 19 database clusters, retaining every question within a sampled cluster and recomputing the equal-question mean. These intervals summarize variation across databases conditional on the recorded eight-continuation pool. Independent-reference squared-error comparisons appear in Appendices H.2 and I.4.

I.6 Response and prediction on the saved ledger pools

This post hoc secondary analysis uses the existing discovery and evaluation pools from Appendix H.6: 48 tasks in four strata, eight blocks per task, four discovery continuations and eight evaluation continuations per block. The consensus target, block library, and direction are frozen before evaluation.

Table 19: **Capacity across continuation budgets.** Equal-question means in pp^2 with 95% database-cluster intervals. The ratio is computed from panel means. All columns use the fresh pool; the original two-continuation estimates are reported in Table 17.

Model	n	$\mathbb{E}_{\mathrm{sub}}[N_n]$	$\mathbb{E}_{\mathrm{sub}}[C_{U,n}]$	Ratio
Qwen	2	605.0 [388.6, 810.4]	369.4 [208.7, 536.9]	1.638
Qwen	4	487.2 [300.6, 671.0]	369.4 [208.7, 536.9]	1.319
Qwen	8	428.3 [255.1, 602.6]	369.4 [208.7, 536.9]	1.159
Gemma	2	250.6 [68.7, 469.6]	170.1 [25.1, 353.5]	1.473
Gemma	4	210.3 [47.8, 410.1]	170.1 [25.1, 353.5]	1.236
Gemma	8	190.2 [36.5, 380.8]	170.1 [25.1, 353.5]	1.118

For each task, we evaluate an exponential tilt of its frozen discovery direction at KL budgets 0, $.01$, $.05$, and $.10$. We retain six constant directions and eleven negative discovery capacity estimates. R denotes agreement with the frozen consensus target, and J denotes joint correctness of the state and action.

Table 20: Ledger response to each task's frozen tilt, in percentage points. R denotes agreement with the frozen target and J joint state-and-action correctness. We assess the same direction in the discovery and evaluation pools; all KL-zero values are exactly zero. Intervals use 20,000 paired whole-task bootstrap draws within the four strata.

Endpoint	KL	Discovery forecast	Evaluation outcome	Discovery − evaluation
R	0.01	2.692 [2.255, 3.124]	1.512 [1.099, 1.940]	1.181 [0.948, 1.414]
R	0.05	5.842 [4.883, 6.791]	3.248 [2.371, 4.161]	2.594 [2.090, 3.100]
R	0.10	8.036 [6.700, 9.354]	4.429 [3.239, 5.669]	3.607 [2.910, 4.301]
J	0.01	1.895 [1.300, 2.459]	0.962 [0.469, 1.448]	0.933 [0.628, 1.229]
J	0.05	4.064 [2.762, 5.289]	2.047 [0.995, 3.079]	2.017 [1.349, 2.658]
J	0.10	5.533 [3.736, 7.222]	2.765 [1.331, 4.165]	2.768 [1.841, 3.652]

At KL $.05$, the average independent gain is 3.248 pp for R and 2.047 pp for J, compared with discovery forecasts of 5.842 and 4.064 pp. The discovery forecast exceeds the independent response at all three positive budgets, with paired intervals above zero (Table 20). Among the 42 nonconstant directions, 37 improve agreement and 32 improve joint correctness at each positive budget. We evaluate both agreement with the learned target and joint task correctness.

Predicting response from discovery measurements. We use four-fold cross-fitting to predict three evaluation targets separately: the R gain and J gain at KL $.05$, and the corrected R capacity. Within each stratum, sorted task identifiers are assigned to folds by index modulo four. Each fit uses ordinary least squares with an intercept and a fixed feature set. We standardize features using training-fold means and scales. The baseline uses discovery agreement q; the other fits add corrected capacity, uncorrected between-block variance, $q(1-q)$, relative resolution, per-token surprisal, or mean sequence surprisal. Relative resolution is the signed corrected capacity divided by $q(1-q)$; its six undefined values are set to zero with a separate indicator. All other additions use one scalar feature.

Response intervals use 20,000 paired whole-task bootstrap draws within the four strata. Prediction intervals use 2,000 whole-task bootstrap refits within each stratum and original fold, keeping every repeated copy of a task in that fold. This preserves the separation of training and evaluation tasks during resampling.

Corrected capacity adds predictive information about agreement gain and evaluation capacity, lowering both squared errors. Errors in predicting joint-correctness gain remain similar for the baseline and corrected-capacity fits. Uncorrected variance gives lower point estimates of prediction error on all three targets. For joint gain, its direct paired comparison with corrected capacity gives an interval below zero: $\mathrm{MSE}_{\mathrm{uncorrected}} - \mathrm{MSE}_{\mathrm{corrected}} = -2.231\ \mathrm{pp}^2\ [-6.921, -0.174]$. Outcome uncertainty $q(1-q)$ gives the lowest point estimate of prediction error for joint gain; its paired interval against

Table 21: Ledger prediction errors from four-fold cross-fitting, with 95% intervals from 2,000 whole-task bootstrap refits. Every entry is the raw mean squared error multiplied by 10^4: the first two columns are in pp^2, whereas the capacity column reports squared error in estimating a variance. Both gain columns use KL 0.05. Each predictor includes discovery agreement q.

Predictor	R gain	J gain	Evaluation capacity
q alone	$5.871\ [4.035, 8.984]$	$18.020\ [10.139, 32.535]$	$14.189\ [7.519, 25.650]$
q + corrected capacity	$3.514\ [2.038, 6.108]$	$18.088\ [8.055, 39.733]$	$6.784\ [3.522, 13.158]$
q + uncorrected variance	$2.965\ [1.719, 4.968]$	$15.857\ [6.850, 33.860]$	$5.915\ [2.788, 11.882]$
$q + q(1-q)$	$5.109\ [3.608, 11.847]$	$10.018\ [6.290, 25.234]$	$12.737\ [8.120, 33.250]$
q + relative resolution	$3.261\ [2.040, 5.720]$	$15.999\ [7.930, 34.657]$	$7.836\ [4.334, 16.204]$
q + per-token surprisal	$6.179\ [4.179, 9.837]$	$18.963\ [10.663, 35.582]$	$14.936\ [7.624, 27.912]$
q + sequence surprisal	$5.925\ [4.090, 9.677]$	$18.476\ [9.643, 36.193]$	$14.506\ [7.770, 27.570]$

Table 22: Paired differences of ledger prediction errors, on the scale of Table 21, with 95% intervals from whole-task refits. Positive values favour the second predictor in each comparison. Baseline denotes q alone; corrected and uncorrected denote q plus corrected capacity and q plus uncorrected variance.

Comparison	R gain	J gain	Evaluation capacity
Baseline − corrected	$2.357\ [0.195, 5.244]$	$-0.068\ [-9.724, 7.298]$	$7.404\ [2.078, 15.075]$
Baseline − uncorrected	$2.906\ [1.232, 5.605]$	$2.163\ [-4.773, 9.175]$	$8.274\ [3.175, 16.416]$
Uncorrected − corrected	$-0.549\ [-1.803, 0.071]$	$-2.231\ [-6.921, -0.174]$	$-0.870\ [-2.766, 0.192]$

the baseline includes zero. Capacity correction isolates variation due to the block; the predictive value of discovery measurements depends on the target outcome.

The all-task response in Table 20 evaluates each task's tilt against uniform block weights; Appendix H.6 evaluates the top-three-per-stratum allocation against random selection.

I.7 Replication on new questions and newly sampled plans

We evaluate forecast accuracy on new questions and newly sampled plan libraries. The new Qwen panel contains 32 previously unused Spider development questions: 14 hard and 18 extra-hard, spanning 17 databases. We apply the original input-based eligibility rules, exclude all previously used questions, and select questions in a fixed deterministic order, cycling across databases. Eligibility is determined by query structure, available test suites, reference-query execution, and prompt length, so the panel is defined from task inputs before generation. The questions are disjoint from the original panel, and the databases come from the same collection.

For each question, we independently generate and freeze $m = 16$ plans. Two discovery continuations per plan define the weights and forecasts in Equations (61) and (62). We freeze these quantities before sampling four independent evaluation continuations per plan, giving 512 plans, 1,024 discovery continuations, and 2,048 evaluation continuations. The model, temperature 1, disabled thinking, prompts, and the expanded-suite correctness criterion match the original experiment. Here $\bar{F}_i^E$ averages four evaluation outcomes. All 64 directions enter (63), including 24 constant directions; 21 questions have at least one nonconstant direction.

We estimate L separately on this panel, squaring each direction's error, then averaging the paired differences within each question and finally across questions. The 95% percentile intervals use 10,000 draws of the 17 database clusters. Each draw includes all questions from each sampled database, with repeated copies when a database is sampled more than once, and gives every question equal weight. The same draws are used for all reported metrics.

Forecast performance. The same-continuation forecast is 5.180 pp, the independent forecast is 3.116 pp, and the independent gain is 3.106 pp (Table 1). The corresponding prediction errors and squared errors appear in Table 23. Independent prediction reduces observed mean squared error from 13.407 to 3.040 pp^2, a 77.3% reduction, with $L = 10.367\ [3.813, 17.709]\ \mathrm{pp}^2$. The gain

Table 23: Prediction errors and covariance on the new-question Qwen panel. Intervals use the same database-cluster draws as the main endpoint. The two signed errors are in pp; all remaining quantities are in pp^2.

Quantity	Estimate	95% interval
Same-continuation error $s - e$	2.074	$[0.992, 3.225]$
Independent error $c - e$	0.010	$[-0.525, 0.525]$
Same-continuation MSE	13.407	$[5.630, 21.874]$
Independent MSE	3.040	$[1.075, 5.399]$
Corrected final capacity $C_{U,4}$	412.3	$[242.2, 593.9]$
Shared-plan covariance $\widehat{\Gamma}_{DF}$	406.8	$[241.8, 581.0]$
Within-continuation covariance W_{DF}	279.9	$[162.8, 407.2]$

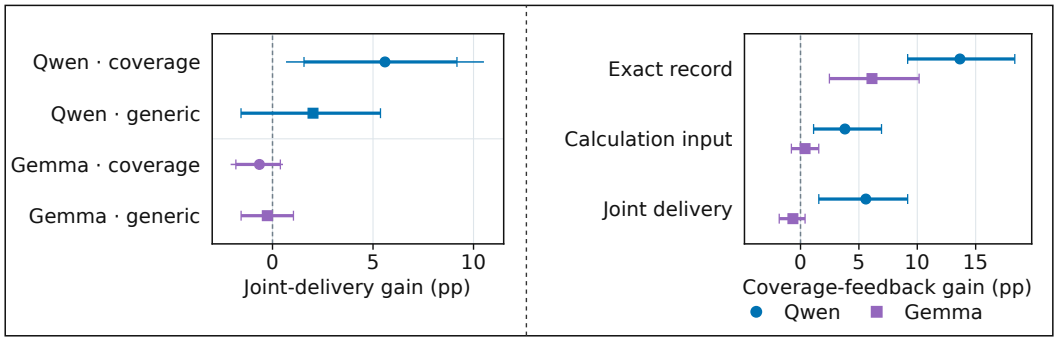

Figure 10: **Feedback value depends on the outcome being improved.** Left: coverage feedback and generic review relative to ordinary continuation. Thick bars show 95% task-resampling intervals; thin bars for coverage feedback also resample repair prefixes. Right: coverage-feedback gains in exact-record correctness, correctness of the effective calculation input, and joint delivery, with 95% task-resampling intervals. Qwen contrasts use the cohort standardization in Appendix H.7; Gemma includes all flagged blocks. Effects are in percentage points.

estimated from four independent continuations provides the reference for these errors and their MSE ratio. The paired loss difference also satisfies $L = \mathbb{E}[(s - c)(s + c - 2e)]$, so the shared squared evaluation-noise term cancels. Removing any one database leaves L positive, ranging from 8.628 to 11.087 pp^2. Together with the original panel, this establishes the forecast advantage on two disjoint sets of questions and separately sampled plan libraries.

Shared plans and continuation dependence. Applying Equations (68) and (70) at $m = 16, n = 4$ gives $T_{DF} = B_{DF} + \frac{3}{4}W_{DF}$ and $\widehat{\Gamma}_{DF} = B_{DF} - \frac{15}{64}W_{DF}$. The shared-plan covariance is 406.8 pp^2 and the within-continuation covariance is 279.9 pp^2, both with intervals above zero. The latter contributes 30.8% of the empirical total covariance through $(3/4)W_{DF}/T_{DF}$. The two panels identify a common mechanism: draft and final outcomes share both a plan-level signal and dependence within a continuation. The coefficients correspond to $n = 4$ here and $n = 8$ on the original panel.

J Revisiting a computation can improve delivery

The preceding interventions keep the continuation policy fixed and change choice weights. Source-checking feedback prompts the model to revisit the computation by reporting discrepancies between its record and the source rows at flagged ledger blocks in a new user turn. This experiment examines a change in K, complementing the fixed-K reweightings. The same gate defines the generic-review and coverage-feedback arms. Qwen uses a prespecified deterministic repair cohort; its contrasts are standardized to the panel's flagged fraction, as defined in Appendix H.7.

Qwen's cohort-standardized coverage contrast is 5.595 pp. Adding it to the ordinary baseline 0.779 gives a standardized level of 0.835. Gemma's all-flagged contrast is -0.651 pp (Figure 10, left).

Table 24: SQL capacity and temporal alignment across input-defined reference conventions. T/DB gives retained tasks/databases; bounds are 95% database-cluster intervals and $\widehat{\rho}_F$ is the point budget fraction. A star marks a panel with zero empirical outcome budget. The extended panel also excludes tasks with the broader reference conventions described below.

Panel (T/DB)	Model	$\widehat{\lambda}_F$	$\widehat{\Gamma}_{DF}$	$\widehat{\rho}_F$
Full (32/19)	Qwen	$0.038\ [0.019, 0.059]$	$0.034\ [0.016, 0.054]$	0.461
Full (32/19)	Gemma	$0.018\ [0.001, 0.039]$	$0.012\ [-0.0003, 0.0287]$	0.540
Top-1 filter (20/15)	Qwen	$0.025\ [0.007, 0.051]$	$0.021\ [0.008, 0.037]$	0.469
Top-1 filter (20/15)	Gemma	$0.014\ [-0.003, 0.042]$	$0.007\ [-0.001, 0.022]$	0.411
Name filter (30/18)	Qwen	$0.041\ [0.020, 0.063]$	$0.036\ [0.018, 0.056]$	0.461
Name filter (30/18)	Gemma	$0.020\ [0.001, 0.042]$	$0.012\ [-0.0004, 0.0306]$	0.540
Key filter (23/14)	Qwen	$0.036\ [0.014, 0.061]$	$0.040\ [0.016, 0.064]$	0.405
Key filter (23/14)	Gemma	$0.026\ [0.002, 0.053]$	$0.016\ [-0.0005, 0.0392]$	0.556
Combined (15/11)	Qwen	$0.019\ [0.007, 0.034]$	$0.024\ [0.009, 0.042]$	0.351
Combined (15/11)	Gemma	$0.019\ [-0.004, 0.054]$	$0.009\ [-0.001, 0.029]$	0.430
Extended (10/8)	Qwen	$0.018\ [0.004, 0.039]$	$0.018\ [0.004, 0.040]$	0.412
Extended (10/8)	Gemma	0^*	0^*	n/a

Interpreting Qwen's standardized level as an all-flagged policy value requires equal mean effects in selected and remaining flagged blocks. Appendix H.7 gives the contrasts and task-cluster and nested-prefix intervals.

Replay distinguishes the recorded evidence, the calculator's effective input, and final delivery (Figure 10, right). Two records can list different evidence while supplying the same values to the calculation. The replay groups partition the reported feedback contrast by the pretreatment record's effect on the reference computation. Qwen gains both when the record changes the reference action and when its differences leave the calculation unchanged; Gemma's response in the latter group has the opposite sign. Measurements of the records and delivery after feedback show which observed quantities change when source feedback prompts new generation (Appendix H.7).

K SQL BRANCH STRUCTURE ACROSS REFERENCE CONVENTIONS

We classify all 32 questions using their reference SQL and rendered schemas. The three input properties are aggregate extrema selected by `LIMIT 1` with unspecified tie-breaking (12 tasks), country exclusion based on the Name field (2), and relevant primary-key descriptions inconsistent with displayed rows or a bundled composite-key declaration (9). Together, these groups contain 17 distinct tasks. Each filter excludes its indicated task group. We use the same retained tasks for both models and retain the original execution labels; Table 24 reports equal-task means with database-cluster intervals.

Qwen's final capacity and draft–final alignment are positive throughout these panels. After the combined structural exclusions, 15 tasks from 11 databases yield $\widehat{\lambda}_F = 0.0192\ [0.0070, 0.0342]$ and $\widehat{\Gamma}_{DF} = 0.0242\ [0.0089, 0.0421]$; the fraction of the outcome's variance explained by plans is $0.351\ [0.211, 0.651]$. The final categorical trace is $0.0692\ [0.0254, 0.1391]$. Excluding the two world questions leaves a capacity of $0.0265\ [0.0128, 0.0445]$ for the suite-mismatch category, extending the categorical finding beyond the illustrated world task.

The extended panel also considers tournament victory versus match victory, identity projection before `INTERSECT`, allocation of votes through state, repeated student enrollment, and numerical ordering of a TEXT column. It retains 10 tasks from eight databases. Qwen's capacity is 0.0175 $[0.0042, 0.0389]$, with cross-checkpoint covariance $0.0181\ [0.0043, 0.0403]$. On Gemma, these ten tasks each have one observed final category and zero empirical outcome budget. Gemma's measured variation is concentrated in the reference-convention groups represented in the larger panels.

Evaluation, flight, country-language and course-arrangement examples contain repeated values in a column described as the sole primary key. For the network tasks, the actual SQLite declaration is `PRIMARY KEY(student_id,friend_id)`. We determine each task's classification and

exclusion status from its question, reference query, and rendered schema. We use all 16 block pairs per retained task and 20,000 database-cluster draws. For normalized intervals, we resample numerators and denominators together. Ratio intervals are reported when every resampled denominator is positive. This sensitivity analysis averages the original paired outcomes over each input-defined task subset; the full benchmark remains the primary panel.

L EXECUTED SQL WITNESSES

We executed a fixed set of saved queries from four questions to connect the categorical measurements to database operations. We check fourteen checkpoint observations using thirteen distinct predictions and their four reference queries. We use the original databases and official expanded-suite databases from the benchmark archives. Re-execution reproduces all fourteen frozen categorical labels. We compare queries on the original databases and on the first suite databases where results differ, then inspect the results, schemas, and relevant rows to connect the measured categories to specific SQL operations.

L.1 A DATABASE IDENTIFIER BECOMES A PLANNING ASSUMPTION

For `car_1_0129`, the question asks for European countries with at least three car manufacturers. Plans 07 and 10 use the literal constraint `Continent = 2`. In the original database, identifier 2 denotes Europe; both their final queries and the reference return France and Germany. In `car_1v5081.sqlite`, identifier 2 has the continent name `4Vz3ILw`. The saved queries return a country named `2`, while the reference, which checks for the continent name Europe, returns an empty result. Thus a planning assumption about a particular database's identifier survives into both final continuations and becomes visible in the expanded execution tests.

Table 25: Selected executed comparisons. Every entry comes from a saved query run against the indicated benchmark database.

Case and database	Reference result	Saved prediction result	Operation explaining the difference
Car, original	France, Germany	France, Germany	Continent ID 2 denotes Europe
Car, suite v5081	Empty	Country `2`	ID 2 names another continent
Network, original	10.428571	Draft: 10.25	Join weights students by qualifying friends
Network, original	10.428571	Revised final: 10.428571	Distinct student IDs restore one contribution per student
World, original	5,451,331,150	Plan 00: 1,192,425,044	Sum of city populations versus country populations

L.2 A REVIEW CHANGES THE AGGREGATION UNIT

In `network_1_0915`, the requested mean is over students who have friends. Both continuations of plan 15 initially average joined student–friend rows and return 10.25 on the original database. The reference uses each qualifying student once and returns 10.428571. One continuation introduces distinct student IDs in its final query and matches all 29 suite databases; the other keeps the join average and returns 10.25. The execution results explain the effect of this SQL change: duplicate friend relationships weight a student's grade repeatedly.

The actual SQLite `Friend` primary key is the composite `(student_id, friend_id)`. The prompt's metadata lists `student_id` as a primary-key field; the database declaration supplies the full two-field key. Inspecting the rows confirms the duplication present in the executed database.

L.3 What the categorical world example distinguishes

For `world_1_0766`, plan 00 sums city populations, producing 1,192,425,044 against the reference country-population total of 5,451,331,150. Plan 01 sums country populations and agrees on the original database. It differs on `world_1v5080.sqlite`: the reference excludes countries by name, while the prediction excludes by country code, and this expanded database contains repeated country names. Their respective totals are 268,945,194,457 and 273,514,308,785. The two categories in Equation (9) therefore distinguish reproducible patterns of agreement with the declared reference across database realizations. All sixteen plans in this example reach the planning marker.

L.4 Tied maxima in the fourth fixed case

The two saved `dog_kennels_0936` finals agree with the reference on the original database. On `dog_kennelsv5080.sqlite`, three owners tie at four treatments. Both the reference and predictions sort by count in descending order and use `LIMIT 1`; the reference and predictions return different tied owners. The counted treatment IDs are non-null and the count expressions agree. The frozen execution-agreement category therefore reflects the queries' different tie selections.

M Independent forecasts for SQL intervention allocation

We use independent gain forecasts to allocate a fixed intervention budget, then examine how accurately they predict gains for the frozen directions, which questions they select, and how these selections improve final correctness. We froze the design, calibration coefficients, and primary comparison before generating the new panel, and froze directions and selections before generating evaluation continuations.

M.1 Panel and sampling protocol

The panel contains 128 Spider development questions whose identities and wordings were absent from the earlier panels: 47 hard and 81 extra-hard questions from 13 databases. We select questions using the input-based eligibility rules of the earlier SQL panels: at least two base tables and a compositional operation, available original databases and expanded test suites, a passing reference-query self-check, and sufficient prompt space. Previously used question identities and identical normalized question texts are excluded. We take questions in a fixed deterministic order, cycling through databases and selecting at most 16 questions per database. These input criteria determine which questions enter the panel. All 13 databases also appear in the earlier panels. The 128 questions form 85 reference-SQL groups after collapsing whitespace and converting SQL text to lowercase within each database; 24 questions share their normalized reference SQL with a historical question on the same database.

For each question, Qwen3.8-27B generates 16 plans, which we then freeze. Each plan receives two independent discovery continuations and eight independent evaluation continuations, yielding 2,048 plans, 4,096 discovery outputs, and 16,384 evaluation outputs. Sampling uses temperature 1, thinking disabled, a 512-token plan cap, and an 8,192-token answer horizon including the retained plan. The prompts and draft/final extraction follow the earlier SQL protocol. Final-query correctness is defined by passing the expanded test suite. Every planned output was generated and scored on its first attempt.

A common direction and budget. Write q for a question, $i \in \{1, \dots, 16\}$ for a plan, and $r \in \{0, 1\}$ for a discovery replica. Let D_{qir} and F_{qir} be the draft and final test-suite correctness labels. A fixed deterministic rule based on question identity chooses a builder replica r_q before generation. Using its draft labels, construct

$$\pi_{qi} = \frac{\exp(\tau_q D_{qir_q})}{\sum_{j=1}^{16} \exp(\tau_q D_{qjr_q})}, \qquad D_{\mathrm{KL}}(\pi_q \| \mu) = 0.05, \qquad a_{qi} = \pi_{qi} - \frac{1}{16}, \tag{73}$$

where $\mu_i = 1/16$ and the nonnegative τ_q is found by bisection using natural logarithms. When the builder's draft labels are constant, $\pi_q = \mu$ exactly. There are 48 such questions and $M = 80$ eligible nonconstant directions. Each selector intervenes on $k = \min(\lfloor 128/4 \rfloor, M) = 32$ questions, with total KL budget $32 \times 0.05 = 1.6$. Unselected questions retain the uniform plan law. All selectors use the same direction for any given question and choose which questions receive it. Direction construction uses oracle test-suite feedback on the discovery drafts.

Selectors. The three gain scores are

$$s_q = \sum_i a_{qi} F_{qir_q}, \qquad c_q = \sum_i a_{qi} F_{qi,1-r_q}, \qquad p_q = \tfrac{1}{2}(s_q + c_q). \tag{74}$$

SAME ranks by s_q, CROSS by c_q, and POOLED by p_q. The three additional selectors use the same two discovery continuations. NAIVE ranks by the population variance across the 16 per-plan mean final scores. ORC-CAPACITY subtracts the finite-replica noise correction from this variance:

$$N_q = \frac{1}{m}\sum_i (\bar{F}_{qi} - \bar{F}_q)^2, \qquad C_q = N_q - \frac{m-1}{mn}\frac{1}{m}\sum_i v_{qi}, \qquad (m,n) = (16,2), \tag{75}$$

where v_{qi} is the unbiased sample variance of the two discovery finals. The ranking uses the raw corrected score. UNCERTAINTY ranks by $\bar{F}_q(1-\bar{F}_q)$, where $\bar{F}_q$ averages all 32 discovery finals. Every deterministic selector takes its top 32 eligible questions; a fixed deterministic ordering of question identities breaks score ties. RANDOM samples 32 eligible questions uniformly, and its expected value is computed analytically using inclusion probability k/M.

Held-out policy value and uncertainty. Let $\bar{F}^E_{qi}$ be the mean of the eight evaluation finals, and define the estimated gain of the frozen direction by $g_q = \sum_i a_{qi}\bar{F}^E_{qi}$. If S_{sq} indicates selection by method s, its gain averaged over the full panel is

$$V_s = \frac{1}{128}\sum_{q=1}^{128} S_{sq} g_q. \tag{76}$$

This is an offline estimate of policy value obtained by weighting independent evaluation outcomes within each frozen plan library. The uniform-plan baseline is $U = \frac{1}{128}\sum_q \frac{1}{16}\sum_i \bar{F}^E_{qi}$; absolute correctness is $U + V_s$. The primary endpoint is $V_{\mathrm{CROSS}} - V_{\mathrm{SAME}}$. All 128 questions, including constant and unselected questions, remain in the denominator.

The primary analysis uses 20,000 bootstrap draws of the 13 database clusters. Each draw includes all questions from each sampled database, repeating them when a database is sampled more than once. We divide the summed gains by the number of included questions, giving each question equal weight. Directions, scores, and selected sets remain frozen. All reported intervals use the same draws and the 2.5th and 97.5th percentiles, holding the evaluation continuations fixed. A prespecified sensitivity analysis reranks the question copies in each draw using the frozen scores and selects $k^* = \min(\lfloor N^*/4 \rfloor, M^*)$. The primary analysis evaluates the originally frozen selections; the sensitivity analysis evaluates the equal-budget selection rule reapplied to each resampled panel. The two sets of intervals quantify uncertainty under these respective procedures.

M.2 Allocation results

CROSS raises final test-suite correctness from the uniform-plan baseline of 49.12% to 51.45%, a gain of 2.33 pp. It exceeds equal-budget random selection by 1.05 $[0.52, 1.49]$ pp. The primary CROSS–SAME difference is 0.22 $[-0.35, 0.78]$ pp; the re-selection sensitivity interval is $[0.07, 0.53]$ pp. Differences from POOLED, NAIVE, ORC-CAPACITY, and UNCERTAINTY range from 0.06 to 0.30 pp; Table 27 gives their intervals. Tables 26 and 27 report every prespecified selector and comparison.

M.3 Forecast accuracy and calibration

Fixed-KL forecasts. We evaluate s_q, c_q, p_q against g_q across all 128 questions, before restricting the comparison to the selector's top-32 questions. Constant directions contribute zero prediction and zero gain. The mean gain across these potential interventions is 3.21 pp. CROSS reduces MSE from 15.61 to 3.15 pp^2, a 79.8% reduction relative to SAME. The paired squared-error difference is 12.46 $[7.67, 18.92]$ pp^2; POOLED minus CROSS is 2.78 $[1.10, 5.20]$ pp^2. SAME overestimates gains by 2.08 $[1.56, 2.72]$ pp, compared with 0.30 $[-0.07, 0.73]$ pp for CROSS. Table 28 reports the full forecast comparison. The 79.8% reduction is measured against the eight-continuation evaluation estimate; the common squared evaluation-noise term cancels in the paired loss difference.

Table 26: All SQL allocation policies use 32 interventions and total KL 1.6. The uniform-plan baseline is 49.12%. Correctness is in percent; gains and their 95% intervals are in pp. RANDOM reports the analytically computed mean for uniform selection among the 80 eligible questions.

Selector	Correctness	Gain	95% interval
SAME	51.23	2.11	[1.22, 2.96]
CROSS	51.45	2.33	[1.57, 3.05]
POOLED	51.33	2.21	[1.48, 2.95]
NAIVE	51.35	2.23	[1.44, 3.07]
ORC-CAPACITY	51.40	2.28	[1.49, 3.09]
UNCERTAINTY	51.16	2.03	[1.38, 2.84]
RANDOM	50.41	1.29	[0.97, 1.60]

Table 27: Prespecified policy comparisons in pp. The primary analysis uses database-cluster intervals with selections held fixed. For the re-selection intervals, we reapply the top-quarter rule within each resampled panel.

Comparison	Estimate	Frozen-selection CI	Re-selection CI
CROSS − SAME (primary)	0.22	[−0.35, 0.78]	[0.07, 0.53]
CROSS − POOLED	0.12	[−0.27, 0.52]	[−0.01, 0.29]
CROSS − NAIVE	0.10	[−0.16, 0.31]	[−0.05, 0.20]
CROSS − ORC-CAPACITY	0.06	[−0.15, 0.22]	[−0.05, 0.15]
CROSS − UNCERTAINTY	0.30	[−0.16, 0.64]	[0.07, 0.48]
CROSS − RANDOM	1.05	[0.52, 1.49]	[0.81, 1.23]

Replication of the two-direction metric. For direct comparison with the earlier panels, we also apply their original rule: each discovery replica defines a direction by doubling the weight of plans whose draft is correct. We square each direction's prediction error, average the two directions within a question, and then average across all questions. This gives $L = 9.60$ $[5.31, 15.97]$ pp^2, following 8.57 $[3.15, 15.53]$ and 10.37 $[3.81, 17.71]$ pp^2 in the earlier Qwen panels. The new panel's mean same-continuation prediction, independent prediction, and realised gain under this rule are respectively 4.91, 3.34, and 3.12 pp. Their biases are 1.79 $[1.34, 2.45]$ and 0.22 $[-0.08, 0.61]$ pp. The two-direction metric uses two fixed-doubling directions per question; the fixed-KL metric uses one direction at KL 0.05. They evaluate different interventions.

Historical affine calibration. To assess how historical calibration affects forecast accuracy, we fit a separate affine map for SAME, CROSS, and POOLED on the two earlier Qwen panels. The fit uses 42 nonconstant directions among 64 historical questions, with corrected labels, and constrains the slope to be nonnegative. The maps, in probability units, are

$$\widetilde{s} = -0.001504 + 0.579223s, \quad \widetilde{c} = 0.019278 + 0.669610c, \quad \widetilde{p} = -0.002788 + 0.798597p.$$

These coefficients are frozen before the new panel is generated. Predictions for new constant directions remain zero. Calibration transforms the forecasts for evaluation; allocation uses the original frozen selector scores.

The fitted maps reduce average bias. After calibration, CROSS achieves 3.55 $[1.60, 5.36]$ pp^2 lower squared error than SAME. The calibrated POOLED–CROSS difference is 0.34 $[-0.44, 1.16]$ pp^2. As a prespecified sensitivity, fitting on the historical labels before the scoring correction gives a SAME–CROSS difference of 3.78 $[1.83, 5.59]$ pp^2 and a POOLED–CROSS difference of 0.34 $[-0.39, 1.15]$ pp^2. Within the calibration inputs, the full recheck of historical scores identifies the same 74 corrected checkpoints used by the frozen fit, including 31 changes in binary correctness. The calibration therefore remains unchanged.

M.4 Selection mechanism and sensitivity analyses

Where the selections differ. We examine the selected questions post hoc to relate forecast differences to allocation. CROSS and SAME share 25 of their 32 selections, with seven further questions

Table 28: Forecast accuracy over all 128 potential interventions. Mean predictions and signed errors are in pp; MSE is in pp^2. Brackets give 95% database-cluster intervals. The corresponding mean held-out gain is 3.21 pp. Calibration coefficients are fitted on historical panels and frozen before this panel is generated.

Forecast	Mean	Bias [95% CI]	MSE [95% CI]
Uncalibrated fixed-KL forecasts			
SAME	5.30	2.08 [1.56, 2.72]	15.61 [9.66, 23.63]
CROSS	3.51	0.30 [−0.07, 0.73]	3.15 [1.72, 5.04]
POOLED	4.41	1.19 [0.84, 1.67]	5.93 [3.31, 9.73]
Historical affine calibration			
SAME	2.97	−0.24 [−0.67, 0.28]	6.67 [4.58, 8.71]
CROSS	3.56	0.34 [−0.003, 0.77]	3.12 [2.33, 4.27]
POOLED	3.34	0.13 [−0.19, 0.58]	3.45 [2.24, 5.09]

Table 29: Post hoc reference-query checks with frozen directions and selections. The allocation difference is CROSS minus SAME in pp; the prediction loss difference is SAME minus CROSS in pp^2 for the fixed-KL forecasts. Each row uses the archived database-cluster bootstrap draws.

Sample and weighting	Allocation difference	Prediction loss difference
All 128 questions	0.219 [−0.349, 0.780]	12.455 [7.674, 18.924]
104 after removing historical overlap	0.297 [−0.275, 0.866]	12.956 [7.937, 20.982]
85 reference-SQL groups, equal weight	0.283 [−0.392, 1.000]	11.920 [6.990, 18.146]

unique to each method. The summed realised gains on these unique selections are 0.555 and 0.274 in probability units; dividing their difference by 128 gives the 0.22 pp policy gap. On the questions selected only by SAME, SAME predicts gains between 0.11 and 0.15, whereas the evaluation gains range from −0.002 to 0.088 and CROSS predicts between −0.01 and 0.07. These selections illustrate how optimism in a continuation's own gain estimate can change allocation.

Contributions to the all-panel gap are +0.21 pp from `dog_kennels` and +0.18 pp from `student_transcripts_tracking`, partly offset by other databases. Leaving out one database at a time gives 13 positive point estimates ranging from 0.01 to 0.39 pp; omitting these two databases separately reduces the gap to 0.01 and 0.05 pp. The largest contributions to the observed policy difference come from these two databases.

Reference-query overlap. For post hoc overlap checks, we normalize reference SQL within each database using the same whitespace and case rules, keep the original directions and selected sets fixed, and reuse the same bootstrap indices. Removing the 24 questions that share a historical reference SQL leaves 104 questions, with 25 frozen selections each for CROSS and SAME. A second check averages within each of the 85 reference-SQL groups and then weights groups equally. This gives equal weight to reference-SQL groups while retaining the frozen selections. Under both weightings, the forecast advantage is positive and its interval lies above zero; Table 29 reports both forecast and allocation contrasts under these weightings.

Other selectors versus random allocation. Additional post hoc comparisons using the same bootstrap draws give gains over RANDOM of 0.83 [0.18, 1.43] pp for SAME, 0.93 [0.44, 1.42] for POOLED, 0.94 [0.40, 1.51] for NAIVE, 0.99 [0.46, 1.53] for ORC-CAPACITY, and 0.75 [0.36, 1.29] for UNCERTAINTY. CROSS–RANDOM is the prespecified comparison in Table 27; these additional comparisons show that several discovery-based scores support useful budget allocation.