TAIL-INFLUENCE SAMPLING FOR CVAR POLICY EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Policies with similar mean returns can differ sharply in rare failures, yet estimating lower-tail conditional value-at-risk (CVaR) accurately can require many costly rollouts. When different conditional components of a stochastic workflow can be queried separately, we ask how to allocate a fixed evaluation budget to estimate a fixed policy's CVaR most accurately. We derive a tail influence for each queryable conditional law that aggregates how its uncertainty affects CVaR across every Bellman reuse. Its variance yields the fixed-design efficiency bound and the oracle Neyman allocation. Tail-Influence Sampling (TIS) estimates these influence scales from a pilot model and reallocates fresh queries toward kernels that matter most for the tail; a visitation-anchored variant protects against pilot underallocation. Under fixed dimension and a positive quantile margin, TIS attains oracle asymptotic variance and first-order MSE including pilot cost, while the anchored variant is within a factor two of the oracle. We also characterize an exact-grid regime in which tail- and mean-optimal allocations coincide. On CliffWalking, TIS reduces MSE by 41% versus learned occupancy and 76% versus complete rollouts at the same charged transition budget. In frozen language-model review workflows, anchored TIS beats an equally regularized mean-influence blend in 23 of 24 MMLU-Pro settings and reaches $2.4 - 3.4\times$ lower MSE than rollouts on six-call FinQA reviews.

1 INTRODUCTION

Average performance can conceal the failures that matter most. A language-model workflow may answer correctly on most runs yet occasionally produce a severe numerical error. Similarly, a robot may complete a task reliably yet rarely enter a state that causes damage. Lower-tail conditional value-at-risk (CVaR) measures average return in a specified worst fraction of runs. Estimating this tail average accurately by Monte Carlo can require many complete runs. Such estimates decide which policy or workflow is safe to deploy. For language-model workflows each run costs several model calls, so the evaluation budget, not the model, often limits how reliably rare severe failures can be measured. In resettable simulators and some generative workflows, however, we can directly sample what happens next from a chosen state and action. This gives the evaluator control over where to collect information. Given a limited evaluation budget, which state–action pairs should we query to estimate a fixed policy's CVaR most accurately?

Fixing the policy does not tell us the transition and reward probabilities, which must be learned from queries. Sampling every state–action pair equally ignores their different roles in producing failure, while sampling according to visitation concentrates the budget on frequently encountered pairs. Neither strategy necessarily identifies where uncertainty about the tail originates. A frequently visited state may have predictable outcomes, while uncertainty at a rarer state can change the estimated severity of failures. This is the central difficulty in evaluating rare failures of language-model agents: runs seldom reach the states that produce the worst outcomes, such as a step at which a tool returned a wrong value, so complete rollouts spend almost all of their budget elsewhere. Effective allocation must account for both outcome variability and its effect on the final tail estimate.

Our answer is a computable notion of tail influence: how uncertainty in a conditional law, or kernel, affects uncertainty in the final CVaR estimate. A kernel receives a high score when its possible outcomes vary substantially and those differences meaningfully change the estimated severity of the

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.

worst runs. Its CVaR effect depends on the continuation model and tail cutoff, and the same kernel can be reused at multiple stages. For example, the same review prompt may recur in a language-model workflow, or a control policy may revisit a state. One observation of this shared behaviour therefore provides information about several parts of the return calculation simultaneously. We combine these effects before measuring their variability to obtain one allocation score per kernel.

Tail-Influence Sampling (TIS) uses a uniform pilot to fit the conditional laws, computes influence scales in the fitted model, and allocates fresh queries in proportion to those scales, with a uniform exploration share. A short pilot can nevertheless miss a rare consequential outcome, causing TIS to underestimate a kernel's importance and give it too few additional queries. Such an error can dominate the final CVaR estimate because the neglected kernel may be precisely where the worst outcomes originate. Anchored TIS averages the influence-based allocation with visitation-based shares, mitigating this failure when visitation remains informative. This follows the established defensive-mixture principle (Hesterberg, 1995; Owen & Zhou, 2000): combine a targeted design with a broader reference design. Under the same conditions, the anchor's asymptotic MSE is at most twice the oracle's.

A tail-specific score is not always needed. On an exact categorical grid, if the probability of a submaximal return is below the tail level, the worst runs contain every submaximal return plus some maximal returns; CVaR then moves exactly with the mean return, and tail influence becomes proportional to mean influence. The divergence between the two scores indicates where tail-specific allocation can pay off: estimated from the pilot, it tells an evaluator whether to allocate by tail influence with the defensive anchor or whether mean-based allocation suffices. The payoff is practical: at matched accuracy, TIS needed 2–4 times fewer queries than complete rollouts on our tabular benchmarks, and the anchor 2–10 times fewer than uniform sampling on held-out FinQA.

Prior distributional-RL inference and efficiency results analyze estimation under specified sampling laws (Zhang et al., 2025; Cheng et al., 2026), adaptive stratification learns allocations for fixed within-stratum quantities (Étoré & Jourdain, 2010; Carpentier et al., 2015), and trajectory designs such as ReVar target mean policy evaluation (Mukherjee et al., 2022). Here the allocation score itself depends on an unknown Bellman continuation model and CVaR cutoff; we derive it and show that learning it together with the allocation recovers oracle first-order MSE, including pilot cost.

In short, our contributions can be stated as follows: *(i) A CVaR-specific allocation signal.* We derive each shared kernel's influence through its Bellman uses. Its variance gives a fixed-design efficiency bound and the scales for classical Neyman allocation (Neyman, 1934). *(ii) Learning the score and allocation.* Under fixed dimension, a positive quantile margin, and suitable pilot and exploration schedules, TIS learns the continuation model, cutoff, and influence scales while attaining oracle asymptotic variance and first-order mean-squared error (MSE), including pilot cost. We also characterize the anchor's asymptotic MSE under the same conditions, within a factor two of the oracle. *(iii) When tail specificity matters.* Models with identical visitation, reward moments, and return laws can need different allocations; on exact grids in a rare-failure regime, tail- and mean-optimal allocations coincide, and their divergence serves as a diagnostic. *(iv) Experiments.* CliffWalking and an 18-case inventory disruption family show gains over learned occupancy, learned mean influence, and rollouts at matched query budgets. On language-model workflows with frozen laws, gains over occupancy and rollouts depend on the generator, budget, and workflow length. Blending controls show that the tail score, not added regularization, drives the anchor's gains; the pilot-estimated divergence selects between tail- and mean-based allocation; and in longer review loops the anchor beats complete rollouts at matched cost.

2 Problem and Estimator

We evaluate a fixed policy in a finite-horizon Markov decision process (MDP) with stationary dynamics. The state and action spaces $\mathcal{S}, \mathcal{A}$ are finite, H is the horizon, and s_0 is the initial state. With $h = H - t$ steps remaining, the policy selects $A_t \sim \pi_h(\cdot \mid S_t)$ and the environment draws $(R_t, S_{t+1}) \sim P_{S_t,A_t}$, with $R_t \in [0,1]$. One transition takes (h, s) to $(h-1, s')$; $h = 0$ ends the run. The policy may depend on h, while $P_{s,a}$ does not. Starting at $S_0 = s_0$, total return is $G_H(s_0) = \sum_{t=0}^{H-1} R_t \in [0, H]$. For a chosen level $\alpha \in (0,1)$, our goal is to estimate $\mathrm{CVaR}_\alpha(G_H(s_0))$: the average return in the worst α-fraction of runs, taking only the required frac-

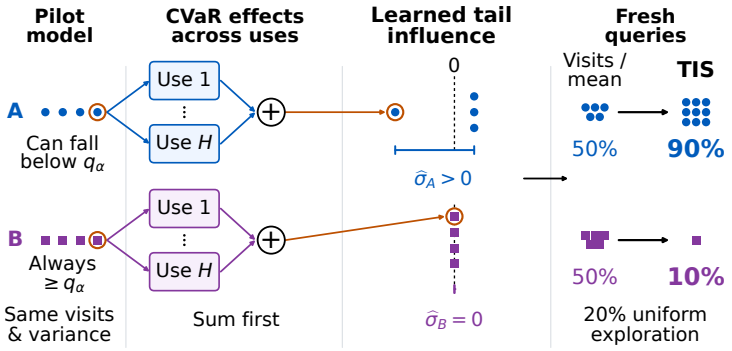


Figure 1: **Equal visitation and reward variability can conceal different information about CVaR.** The same kernel can be reused across many stages (Use 1, . . ., Use H). A and B are visited equally and have identical reward means and variances, but only A's returns can fall below the CVaR cutoff q_α. TIS sums each pilot draw's first-order CVaR effects across all uses, then measures their spread $\widehat{\sigma}_g$ across draws; orange rings track one draw. With 20% uniform exploration, the illustrated scores give nominal fresh-query shares of 90%/10% for A/B, versus 50%/50% for visitation or mean influence (before minimum counts and rounding).

tion of probability mass at the cutoff. Smaller α emphasizes rarer outcomes. Appendix B.2 gives the formal definition.

What can be sampled? A query group $g = (s, a) \in \mathcal{G}$ is one independently queryable conditional law, or kernel. One query returns $W_g = (R, S') \sim P_g$; reward and next state may be dependent. Queries are i.i.d. within a group and independent across groups. For example, an evaluator can reconstruct a review prompt from a specified answer and confidence, then sample its response without reaching that state by rollout. We retain a fixed set of states and $G = |\mathcal{G}|$ groups, closed under all possible continuations. Allocating n_g queries to each group costs $N = \sum_g n_g$; a rollout costs one query per transition. The same fitted law $\widehat{P}_g$ is used at every remaining horizon where that group occurs. Thus $g = (s, a)$ identifies what we sample, while (h, s) identifies where we evaluate its consequences. The same next state can have different continuation returns with one or three steps left; these differences matter for the allocation score.

How do samples give a CVaR estimate? We first estimate the return distribution from each state with h steps left. Following categorical distributional RL (Bellemare et al., 2017), we store probabilities on a fixed return grid $\mathcal{Z} = \{0 = z_0 < \cdots < z_{K-1} = H\}$ with maximum gap Δ. The vector $p^*_{h,s}$ contains these probabilities: $p^*_{h,s,k}$ is the mass at return z_k. The indices mean remaining steps (h), current state (s), and return value (k).

A Bellman update combines the immediate reward with the distribution of future returns. The matrix $Q(R)$ adds R to each grid value and projects the result onto the grid, splitting mass between neighboring points and clipping outside the endpoints (Bellemare et al., 2017; Rowland et al., 2018). Averaging over actions and next transitions gives

$$p^*_{h,s} = \sum_a \pi_h(a \mid s)\, \mathbb{E}_{(R,S')\sim P_{s,a}}\left[Q(R)p^*_{h-1,S'}\right], \qquad p^*_{0,s} = e_0, \tag{1}$$

where e_0 puts all mass at zero. Replace each expectation by a group sample average and evaluate successively for $h = 1, \ldots, H$. CVaR of the estimated root probabilities is $\widehat{C}_N$; $C_{\alpha,K}$ uses the population probabilities (Appendix A). Allocation controls sampling error around $C_{\alpha,K}$. The grid introduces a separate approximation error of at most $H\Delta$ relative to true-return CVaR (Appendix E.1). All limits keep the horizon, state and action sets, grid, tail level, and retained groups fixed. Stars denote population quantities, hats estimates, and (0) pilot estimates.

3 From One Observation to an Optimal Allocation

More queries to a group help only if uncertainty in that group affects the final CVaR estimate. We derive that effect in three steps: write CVaR using an expected shortfall, trace a transition error to that shortfall, then allocate samples according to the variability of the resulting effects.

Read CVaR as a shortfall. For a grid threshold $q \in \mathcal{Z}$, the shortfall $(q - Z)_+$ is zero above q and measures the distance below it. Let $U_h^*(s, q) = \mathbb{E}[(q - Z_h(s))_+]$, where $Z_h(s)$ has the grid probabilities $p_{h,s}^*$. At the root α-quantile q_α, the standard shortfall identity gives

$$C_{\alpha,K} = q_\alpha - \frac{U_H^*(s_0, q_\alpha)}{\alpha}. \tag{2}$$

For probabilities $(.04, .08, .88)$ at returns $(0, .5, 1)$, the worst 10% contains all zero returns and enough .5 returns to average .3. Here $q_\alpha = .5$ and the mean shortfall is $.04 \times .5 = .02$; the formula gives $.5 - .02/.1 = .3$. We assume a positive quantile margin: the tail cutoff lies strictly inside the probability mass at one grid value. In the example, $.04 < .1 < .12$, so small probability errors leave $q_\alpha = .5$ unchanged. With that threshold unchanged, CVaR error is shortfall error scaled by $-1/\alpha$. This lets us focus on one expected shortfall. Appendix B.2 defines the margin and explains why a zero margin can invalidate the Gaussian limit.

Trace a transition error to CVaR. After observing reward R, the remaining shortfall threshold is $q - R$. The sampled target $T_h(W; U)$ therefore uses $U_{h-1}(S', q - R)$, interpolating between grid thresholds and taking zero at nonpositive thresholds; at $h = 1$ it is $(q - R)_+$. Categorical projection preserves these shortfalls at grid thresholds (Lemma 1), so this is the same return calculation in more convenient coordinates.

Write all (h, s, q) shortfalls as the vector U^*. To learn how errors in them affect the initial state's shortfall, compute sensitivity weights r, also called adjoint weights. The two Bellman passes can be written as

$$U^* = \mathbf{b} + MU^*, \qquad (I - M)^\top r = e_{x_0}, \quad x_0 = (H, s_0, q_\alpha). \tag{3}$$

Here $\mathbf{b}$ contains terminal contributions, M carries continuation weights, and e_{x_0} selects the root shortfall. The first equation computes shortfalls; the second assigns a weight to each coordinate according to its effect on the root. Both are computed by passes through the layers, without a dense matrix inverse.

For one observation W from group g, let $\mathcal{T}_g(W; U^*)$ collect its updates wherever that group is used, including policy weights. Subtracting the expected updates gives the one-draw error $\Xi_g(W) = \mathcal{T}_g(W; U^*) - \mathbb{E}\mathcal{T}_g(W; U^*)$. Weighting by r translates this error to the root, and $-1/\alpha$ converts it to CVaR error:

$$\phi_g(W) = -\frac{1}{\alpha} r^\top \Xi_g(W), \qquad \sigma_g^2 = \mathbb{E}\big[\phi_g(W)^2\big]. \tag{4}$$

Thus ϕ_g is one draw's first-order effect on CVaR, and σ_g measures how much that effect varies across draws. A shared kernel's observation affects several stages together. We sum those effects before taking their variance, retaining the cross-stage covariance of the same draw (Equation 16).

Figure 1 illustrates this calculation; Appendix B.2 gives a numerical example and its covariance calculation.

Allocate where more samples reduce error. Averaging n_g independent observations reduces group g's leading variance contribution to σ_g^2/n_g. Contributions from independently sampled groups add. The next theorem formalizes the resulting prediction: CVaR MSE is approximately $\sum_g \sigma_g^2/n_g = V(w)/N$, where w_g is the group's budget share.

Theorem 1 (Allocation-dependent limit)**.** *Under the fixed-dimensional model and positive margin above, let the counts be deterministic with $n_g/N \to w_g > 0$. Then*

$$\sqrt{N}\Big(\widehat{C}_N - C_{\alpha,K}\Big) \Rightarrow \mathcal{N}\big(0, V(w)\big), \qquad N\mathbb{E}\Big[(\widehat{C}_N - C_{\alpha,K})^2\Big] \to V(w), \quad V(w) = \sum_g \frac{\sigma_g^2}{w_g}. \tag{5}$$

Increasing a high-influence group's share reduces its contribution to error. The proof also controls the nonlinear Bellman remainder and wrong-quantile probability in normalized MSE (Appendix C.1).

Theorem 2 (Fixed-design efficiency)**.** *For each such allocation and population law satisfying the positive quantile margin, $V(w)$ is the local semiparametric efficiency bound for $C_{\alpha,K}$ in the product model of unrestricted group laws on their fixed declared outcome spaces, relative to its differentiable-in-quadratic-mean (DQM) tangent space. The empirical categorical Bellman estimator is regular under these local submodels and attains the bound.*

For a fixed budget split, this is the smallest leading variance among regular estimators using the stated conditional samples (Appendix C.2). We can now choose the split itself. When $\sum_g \sigma_g > 0$, minimizing $V(w)$ gives the classical Neyman rule

$$w_g^* = \frac{\sigma_g}{\sum_j \sigma_j}, \qquad V^* = \left(\sum_g \sigma_g\right)^2. \tag{6}$$

A group with twice the influence scale receives twice the oracle share. If some scales vanish, the optimum over positive shares is approached by letting their exploration shares tend to zero. The next example shows why ordinary visitation and reward variability cannot replace this tail-specific score. In a one-step model with equally visited kernels, one returns 0 with probability $.1$ and $5/9$ otherwise; each other kernel returns $1/3$ or $2/3$ with equal probability. Their first two moments match, but only the first has random shortfall below the CVaR threshold $1/3$. Appendix C.4 gives the proof and fixed-root-law family.

Proposition 1 (Equal visitation and reward variance can hide tail influence)**.** *For every fixed $G \geq 2$, there is a one-step categorical model with a uniform policy, conditional mean $1/2$ and variance $1/36$ in every kernel, and a positive margin at $\alpha = .1$, for which*

$$V_{\text{occupancy}} = V_{\text{mean}} = \frac{1}{G}, \qquad V^* = \frac{1}{G^2}.$$

These are the coefficients of $1/N$ in asymptotic CVaR MSE; V_{mean} uses the allocation optimal for estimating the mean return. There is also a family with these same visitation probabilities, conditional moments, and entire root return law in which the occupancy-to-oracle ratio ranges from 1 to G.

With ten kernels, the oracle thus has one tenth of uniform's leading MSE.

4 Tail-Influence Sampling

The oracle shares require the unknown transition laws and their effects on future returns. TIS learns both from a pilot. Algorithm 1 (Appendix D) draws $m = m_N$ samples per group, fits a provisional model, and evaluates it by Equation 1. This gives the pilot root quantile; Equation 3 then gives the shortfalls $\widehat{U}^{(0)}$ and sensitivity weights $\widehat{r}^{(0)}$ needed to score each draw.

Score each draw, then measure variability. For pilot outcome $W_{g,i}$, compute

$$\widehat{d}_{g,i} = -\frac{1}{\alpha}\widehat{r}^{(0)\top}\mathcal{T}_g(W_{g,i}; \widehat{U}^{(0)}), \qquad \widehat{\sigma}_g = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\widehat{d}_{g,i} - \bar{d}_g)^2}, \tag{7}$$

Here $\bar{d}_g$ is the group's average score. Each score combines all stage effects of a draw; larger estimated scales receive more queries:

$$\widehat{w}_g = (1 - \lambda_N)\frac{\widehat{\sigma}_g}{\sum_j \widehat{\sigma}_j} + \frac{\lambda_N}{G}, \qquad 0 < \lambda_N < 1. \tag{8}$$

When all scales vanish, use uniform sampling. For scales 1 and 3, the shares before the floor are $1/4$ and $3/4$. The floor reserves samples for groups the pilot may have underestimated.

After two main samples per group, largest-remainder rounding spends the remaining budget exactly. Only fresh main samples enter the final estimate. As N grows, a larger pilot can consume a shrinking budget fraction. Under the following schedules, TIS attains oracle leading MSE, including pilot cost.

Theorem 3 (Oracle adaptation)**.** *Under the same model and margin, suppose* $\sum_g \sigma_g > 0$*. If* $m_N \to \infty$*,* $Gm_N = o(N)$*,* $\lambda_N \to 0$*,* $\sqrt{N}\lambda_N \to \infty$*, and* $\log(1/\lambda_N) = o(m_N)$*, then*

$$\sqrt{N}\Big(\widehat{C}_N^{\mathrm{TIS}} - C_{\alpha,K}\Big) \Rightarrow \mathcal{N}(0, V^*), \qquad N\mathbb{E}\left[\Big(\widehat{C}_N^{\mathrm{TIS}} - C_{\alpha,K}\Big)^2\right] \to V^*. \tag{9}$$

A total pilot of order $N^{2/3}$ *and floor* $N^{-1/4}$ *suffice for fixed* G*. Individual zero-influence groups are allowed.*

The theorem controls learning the score through the unknown continuation model and quantile, as well as learning its variance and allocation. The proof handles rare underallocating pilots, the nonlinear Bellman remainder, and wrong quantile atoms (Appendix D.3). The limit is pointwise in the fixed model, not uniform over increasingly rare events or shrinking margins.

A defensive anchor for finite pilots. A small pilot can miss a consequential outcome and give its group too few main samples. Local stability near the oracle (Proposition 3) does not control that failure. Following the established defensive-mixture principle (Hesterberg, 1995; Owen & Zhou, 2000), we combine tail targeting with pilot-estimated visitation. This can protect kernels that the pilot recognizes as frequently reached even when their tail influence is underestimated. Let $\widehat{o}_{s,a} = \sum_h \widehat{\mu}_h(s)\pi_h(a \mid s)$ be the pilot-model expected visit count. Form a floored occupancy design from these scores, using the same pilot and floor as the influence design. Anchored TIS averages the two:

$$w^{\mathrm{anc}} = \tfrac{1}{2}(w^{\mathrm{inf}} + w^{\mathrm{occ}}). \tag{10}$$

The influence component keeps its uniform fallback. Use these shares in Algorithm 1. The mixture keeps at least half of either component's share for every group. Before rounding, its leading variance is therefore at most twice that of the better component (Proposition 4, Appendix D.4). Both components can still underallocate the same group. Under the assumptions and schedules of Theorem 3, the anchor's asymptotic MSE constant satisfies $V^* \le V_{\mathrm{anc}} \le 2V^*$ (Corollary 1, Appendix D.4). This does not guarantee lower finite-budget MSE. Plain TIS is the efficient limit when pilots estimate the scales reliably; the anchor pays at most a factor two in the constant for protection against pilots that miss rare outcomes. Pilot underallocation in the initial language-model experiments motivated this anchor. We fixed its design before collecting held-out FinQA numerical-review data (Section 5).

When can a tail-specific score help? Learning an allocation must repay its pilot. Let ρ be the pilot's fraction of the budget, $A_v = V(v)/V^*$ the oracle's advantage over a fixed design v that spends all N queries, and $D(w^*\|w) = \sum_g (w_g^* - w_g)^2/w_g$ the error of the realized main-sample shares w relative to the oracle shares w^* in Equation 6. For the leading error term, learning beats v exactly when

$$1 + \mathbb{E}D(w^*\|w) < (1-\rho)A_v \tag{11}$$

(Appendix D.5). Large allocation advantages, small pilots, and accurate shares favor learning. Because D divides by w_g, a starved group is especially costly; the anchor guards against exactly this. A further limit is structural.

Proposition 2 (Rare failures make CVaR a mean)**.** *Suppose the categorical grid contains every partial return permitted by the fixed declared outcome spaces, so the recursion is exact throughout the product model. Let* $X := G_H(s_0)$ *and let* $x^\star$ *be the maximum return permitted by those outcome spaces. Write* $\pi = \Pr(X < x^\star)$*. If* $\pi < \alpha$*, then*

$$\mathrm{CVaR}_\alpha(X) = \frac{\mathbb{E}[X] - (1-\alpha)x^\star}{\alpha},$$

and locally every kernel's categorical-CVaR influence is $1/\alpha$ *times its ordinary mean-return influence. Hence the tail- and mean-optimal Neyman allocations coincide whenever the influence scales are not all zero; if they are all zero, every allocation has zero first-order variance.*

The worst α-fraction of runs then contains every submaximal return plus enough maximal returns to fill the tail, so CVaR moves exactly with the mean. A mean score is also easier to learn, since it does not depend on a quantile. The total-variation distance $\frac{1}{2}\sum_g |p_g - m_g|$ between normalized tail and mean influence shares (using the uniform vector for an all-zero score) is zero under the proposition and serves as a diagnostic (Appendix E.6).

Table 1: **TIS gains under concentrated tail influence; elsewhere its pilot does not repay.** MSE/uniform MSE ($G = 10, \alpha = .1$); columns: queries/kernel incl. pilots. Uniform=known occupancy; occupancy+pilot isolates pilot cost; Oracle+floor=population scores. Bold: sample-only column minima (point estimates). Table 3: absolute errors/SEs.

	$t = 0$ (equal)		$t = \frac{1}{2}$		$t = 1$ (one active)	
Method	100	400	100	400	100	400
Uniform	**1.00**	1.00	1.00	**1.00**	1.00	1.00
Occupancy + pilot	1.60	1.31	1.67	1.39	1.69	1.31
Learned mean	1.63	1.28	1.71	1.39	1.87	1.34
Complete rollout	1.01	**0.95**	**0.99**	1.07	1.11	1.18
TIS	5.42	4.30	3.62	3.15	**0.22**	**0.16**
anchored TIS	2.04	1.52	1.66	1.17	0.37	0.27
Oracle + floor	1.00	1.00	0.75	0.77	0.13	0.12

5 EXPERIMENTS

We ask five questions: **Q1** Can visitation and mean-return scores miss tail influence? **Q2** Does learning this signal improve matched-budget accuracy? **Q3** Why can pilots fail, and does anchoring help? **Q4** Does anchoring transfer to held-out numerical-review workflows? **Q5** Is the gain tail-score-specific, and does the divergence diagnostic predict where? Theory predicts three regimes: plain TIS should win with concentrated tail influence and reliable pilots (Q1–Q2); the defensive anchor should matter when pilots miss rare outcomes (Q3–Q4); and, on this diagnostic's exact closed grids, no tail-specific gain should appear when submaximal returns are rarer than the tail level (Q5).

Compared methods. Uniform splits queries evenly. Learned occupancy/mean use pilot-estimated visit counts/mean-return influence, respectively, with TIS's pilot, floor, and CVaR estimator. Complete rollouts use whole trajectories, not conditional-query methods' direct kernel access; comparisons are cost-matched, not equal-access. Oracle+floor: exact tail-influence scales, no pilot. Budgets include discarded pilots and all rollout transitions; MSE uses exact closed-grid targets (Appendix E.2).

Q1. Controlled separation. Table 1 varies tail-influence concentration while preserving visitation, conditional moments, and root return law. At $t = 1$, plain TIS is .16–.22 of uniform MSE (oracle .12–.13); at $t = 0$ (uniform optimal), it is 4.3–5.4× uniform MSE. The intermediate case likewise does not repay the pilot (Appendix D.5).

Q2. Tabular benchmarks. Seasonal inventory ($H = 8$, 41 blocks, $\alpha = .1$) uses the prespecified pilot/floor; at 1,200 queries/block, MSE/uniform is .657 for TIS, .918 for learned occupancy, .866 for learned mean influence, and 2.22 for complete rollouts at matched transition cost (Figure 2(b), Table 4). Slippery CliffWalking ($H = 20$, $\alpha = .1$, 149 stationary kernels) tests layer reuse (Figure 2(a)). At 400 queries/kernel, TIS lowers MSE by 40.9% vs. learned occupancy, 18.3% vs. learned mean influence, 76.3% vs. complete rollouts at matched transition cost, and 31.3% vs. population occupancy (Table 6). FrozenLake/rainy Taxi add shared-kernel checks (Appendix E.2.5). Most gains come from pooling reused kernels; covariance terms matter little numerically here (Appendix E.2.4). To test breadth, we prespecified an 18-case family before simulation: disruption probabilities $\{.01, .04, .12\}$, disruption losses 1–3 units and two fixed ordering policies; all 18 reported (Appendix E.4). At 1,200 queries/block, both TIS and the anchor have resolved lower MSE vs. learned occupancy/rollouts in all 18 and vs. learned mean in 17 (median MSE/uniform: .61 TIS, .69 anchored, .93 learned occupancy, .84 learned mean, 2.56 rollouts). At matched RMSE, TIS's query ratios (learned occupancy/rollouts) are .65–.84/.32–.45 on CliffWalking and .55–.72/.26–.30 on inventory; these retrospective interpolations include pilots (Appendix E.2.7).

Q3. Pilot reliability in language-model workflows. Each of 50 fixed MMLU-Pro questions (Wang et al., 2024) is a separate workflow. State $s = (j, c)$ records latest answer $j \in \{1, \ldots, 10\}$ and confidence $c \in \{.1, \ldots, .9\}$. Fixed policy solves first, then selects `reconsider`, `challenge`, or `verify` by confidence band. Prompts include question/current pair/prescribed action, but no history or stage index. Root plus 10×9 pairs yield 91 directly queryable kernels reused across stages.

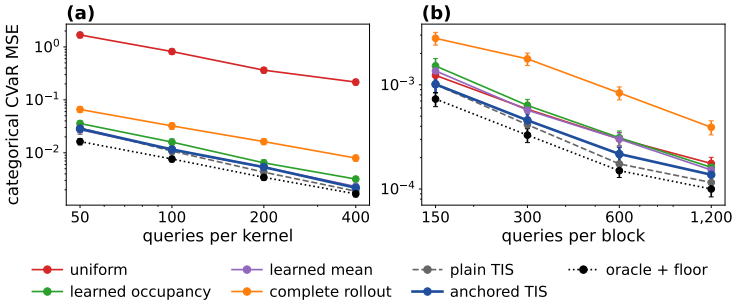


Figure 2: **At largest budgets, TIS MSE is lower than learned occupancy, learned mean, and rollouts.** CVaR MSE vs. charged queries/kernel in (a) CliffWalking and /block in (b) inventory. Conditional-query methods have direct kernel access; rollouts use trajectories at same charged transition budget. Oracle+floor=population scores; bars=1.96 Monte Carlo SEs.

Runs use $H \in \{2, 4, 6\}$ calls. Responses earn normalized Brier utility vs. correct option (Equation 65); for six generators, we estimate each question's worst-10% CVaR against exact closed-grid targets (Appendix E.2.6). Queries sample frozen LLM laws; live GPU/API cost is out of scope.

At $H = 6$, 400 queries/kernel, plain TIS exceeds uniform MSE for Qwen3-4B (1.38) and GLM-4-32B (1.77) (Table 7). Occupancy has lower observed MSE than plain TIS for all six generators. The anchor mitigates both failures, reaching .056–.093 of uniform MSE and the lowest MSE for 5/6 generators; occupancy and complete rollouts remain strong. Kernel reuse matters: fitting recurring-prompt copies separately by step matches shared TIS at $H = 2$ but has 4–10$\times$ its MSE at $H = 4, 6$ (Table 9). The MMLU replay's 90th-percentile realized/oracle variance ratio falls from 136 (plain TIS) to 5.8 (anchor) (Appendix Figure 5(a)): starved groups drive failures, as Equation 11 predicts. Using population influences and realized counts, the variance formula predicts observed MSE without fitted constants (median log-ratios $-.001$ for MMLU and $-.006$ for FinQA Phi; retrospective check, Appendix E.10).

Q4. Held-out financial numerical review. FinQA has numerical questions over real financial reports (Chen et al., 2021). We compare ordinary review and explicit unit-and-sign audit. Each workflow selects one of eight frozen candidate solutions, then reviews twice ($H = 3$) under the MMLU confidence-band policy. Root plus 8×9 candidate–confidence pairs yield 73 queryable kernels; review templates alter transition laws. Terminal utility $u(s) \in [0, 1]$ falls with relative numerical error against gold answer; invalid candidates earn zero. With zero root utility, rewards $R = [u(s') - u(s) + 1]/2$ sum to $G_H = [H + u(S_H)]/2$, so only final-answer utility matters. A 20-question development phase fixed the screen, generators, workflows, seeds, anchored score, and floor before held-out calibration on 50 fresh screened questions from 312 scanned (Appendix E.3). At 100–200 queries/kernel, anchored TIS is .089–.100 of uniform MSE on Qwen3-4B and .216–.327 on Phi-4-mini (Figure 4(a,b)); plain TIS is unresolved vs. uniform in 7/8 cells. At 400/800 queries, the anchor beats occupancy and rollouts in both Phi workflows (MSE ratios .81–.86 and .68–.88, respectively); all 8 contrasts resolve. Occupancy and rollouts remain better on Qwen (Appendix E.3), whose near-deterministic questions fall in the rare-failure regime of Proposition 2. With six vs. three calls (same frozen kernels), the Phi anchor's MSE is .29–.64 of rollouts' at every budget in both workflows, all resolved (Table 17).

Q5. Is the tail score itself what helps? Two declared controls test whether anchor gains come from the tail score: each blends the same floored occupancy shares with uniform or learned mean-influence shares and pays the same pilot (Appendix E.5). Anchor MSE is resolved lower vs. uniform blend in 160/166 settings and higher in none (Table 2), so gains are not a regularization artifact. Vs. mean blend, results follow Proposition 2. FinQA, including calculator-fault variants (Appendix E.7): median tail–mean distance is .000; mean blend is as good or better. MMLU-Pro (more frequent failures; distance .08–.33): anchor MSE is resolved lower vs. mean blend in 23/24 confident-error cells (this utility makes confident mistakes nearly worthless) and 14/15 Brier cells at distance $\geq .24$, versus 2/12 below .18. With pre-simulation predictions, the rule held in all 3 decisive new settings and 6/8 longer-loop settings; the two misses (Qwen3-4B at $H = 8, 10$, distance .26) favored the anchor without resolving (Appendices E.6 and E.8). In practice, use pilot-estimated distance: anchor

Table 2: **Anchor vs. equally regularized blends: never resolved worse than uniform; mean-blend gains depend on domain.** Counts of resolved ($|z| \geq 2$) lower (higher) anchor MSE vs. each blend. Inventory: 18 cases at 1,200 queries/block; FinQA: 2 generators $\times$ 2 workflows $\times$ 4 budgets.

Domain	vs. occupancy+uniform	vs. occupancy+mean
Inventory disruption family	18/18 (0)	17/18 (0)
FinQA review (held-out)	16/16 (0)	0/16 (15)
MMLU-Pro, Brier utility (6 generators $\times$ 3 budgets)	18/18 (0)	11/18 (0)
MMLU-Pro high-stakes panel (3 generators $\times$ 3 budgets)	9/9 (0)	5/9 (2)
MMLU-Pro, confident-error utility (6 generators)	24/24 (0)	23/24 (0)
FinQA with calculator faults (ordinary review)	15/18 (0)	0/18 (12)
FinQA with calculator faults (unit check)	15/18 (0)	0/18 (10)
Prospective divergence test (5 new settings $\times$ 3 budgets)	15/15 (0)	11/15 (1)
MMLU-Pro longer loops ($H = 8, 10$; 3 generators $\times$ 3 budgets)	18/18 (0)	12/18 (0)
FinQA longer reviews ($H = 6$; 2 generators $\times$ 2 workflows $\times$ 3 budgets)	12/12 (0)	0/12 (9)

if pilot median is $\geq .21$, otherwise mean blend; this picked the better or tied design in 141/148 cells (Appendix E.9).

6 Related Work

From distributional inference to query design. Categorical distributional RL supplies the representation and projection (Bellemare et al., 2017; Rowland et al., 2018). Zhang et al. (2025) derive return-law and functional limits under specified, possibly nonuniform, sampling laws. Quantile-based evaluation also has semiparametric efficiency guarantees (Cheng et al., 2026). Shortfall identities and efficiency theory are established (Rockafellar & Uryasev, 2000; van der Vaart, 1998); our addition is the computable, learnable Bellman allocation signal.

Adaptive allocation with a learned Bellman model. Adaptive stratified sampling learns the variances required by Neyman allocation (Neyman, 1934; Étoré & Jourdain, 2010; Carpentier et al., 2015). Theorem 3 controls learning our Bellman-dependent score and its allocation, including pilot cost. Small-pilot failures and defensive mixtures motivate the anchor (Cai & Rafi, 2022; Hesterberg, 1995; Owen & Zhou, 2000); SaVeR targets trajectory design for mean evaluation (Mukherjee et al., 2024). Risk-sensitive control changes the policy (Tamar et al., 2015; Bäuerle & Jaśkiewicz, 2024); with generative access, Deng et al. (2025) study sample complexity for iterated-CVaR policy optimization. We instead fix the policy and optimize conditional-query allocation for estimating its tail functional. Appendix F compares access models and guarantees.

7 Discussion

Allocating queries to estimate CVaR is harder than allocating them for a mean: a kernel's value depends on an unknown continuation model and tail cutoff, and one draw of a reused kernel affects several Bellman stages at once. Our contribution is a computable, learnable CVaR allocation signal for reused conditional laws, with oracle first-order MSE under the stated assumptions, a defensive variant within a factor two of the oracle, and, on exact grids, a condition under which mean influence suffices.

TIS is most promising when conditional access is feasible, tail influence differs meaningfully from visitation or mean influence, and pilots estimate that difference reliably enough to repay their cost. By Equation 11, the failures reflect either little opportunity ($t = 0$ in Table 1) or opportunity the pilot cannot learn (plain TIS in deep loops; FinQA-Qwen, whose rare errors pilots seldom see). In practice, the pilot's tail–mean divergence selects the allocation, at the same computation as mean influence and more than occupancy or rollouts (Table 13). The setting arises wherever an evaluator can restart from a chosen state: auditing rare severe errors of language-model review and agent loops before deployment, where each query is a model call and recurring prompts make kernel reuse the norm, or disruption losses of inventory and maintenance policies in simulators. Tool failures and multi-turn safety evaluation are natural next targets. The guarantees require fixed-dimensional categorical models, a positive quantile margin, independent queries, and the stated pilot and floor schedules. Finite-budget MSE guarantees remain open.

AI USE STATEMENT

We used generative AI tools to edit prose for readability and assist with software coding. All AI-assisted material was reviewed by the authors, who take responsibility for the final text and code.

ETHICS STATEMENT

The experiments use public benchmark questions, simulator environments, and model-generated outputs; they do not collect personal data or involve human subjects. The language-model results may inherit limitations or biases from the evaluated models and benchmark, so we restrict the conclusions to the fixed panel, policy, and utility studied here and report the observed negative result.

REPRODUCIBILITY STATEMENT

The main text specifies the estimator, assumptions, and allocation algorithm. Appendices A–D give the main derivations and proofs, including the anchor's leading-variance bound; Appendix E adds the representation-error and tail–mean-coincidence arguments and documents the experimental protocols, uncertainty calculations, and diagnostics. The supplementary code ZIP contains implementations, pinned dependencies, experiment configurations and seeds, frozen question panels and conditional kernels, recorded results, and provenance manifests. Its README gives setup and verification instructions. The language-model allocation experiments can be rerun on CPU from the frozen kernels without generating new model responses; protocol files record model revisions and sampling settings. The code is also available at https://anonymous.4open.science/r/TIS-F764/README.md.

REFERENCES

Nicole Bäuerle and Anna Jaśkiewicz. Markov decision processes with risk-sensitive criteria: An overview. *Mathematical Methods of Operations Research*, 99(1):141–178, 2024. doi: 10.1007/s00186-024-00857-0.

Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011. doi: 10.1007/s00186-011-0367-0.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458. PMLR, 2017.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Yong Cai and Ahnaf Rafi. On the performance of the Neyman allocation with small pilots. *arXiv preprint arXiv:2206.04643*, 2022. Version 4, revised June 2024.

Alexandra Carpentier, Rémi Munos, and András Antos. Adaptive strategy for stratified monte carlo sampling. *Journal of Machine Learning Research*, 16(68):2231–2271, 2015.

Yash Chandak, Scott Niekum, Bruno Castro da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S. Thomas. Universal off-policy evaluation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27475–27490, 2021.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.300.

Zijie Cheng, Yang Peng, and Zhihua Zhang. Statistical efficiency and inference of quantile distributional reinforcement learning. *arXiv preprint arXiv:2607.08444*, 2026.

Jessica Dai, Paula Gradu, and Christopher Harshaw. Clip-OGD: An experimental design for adaptive Neyman allocation in sequential experiments. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32235–32269, 2023.

Zilong Deng, Simon Khan, and Shaofeng Zou. Near-optimal sample complexity for iterated CVaR reinforcement learning with a generative model. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3907–3915. PMLR, 2025.

Connor Douglas, Joel Persson, and Foster Provost. Logging policy design for off-policy evaluation. *arXiv preprint arXiv:2605.15108*, 2026.

Pierre Étoré and Benjamin Jourdain. Adaptive optimal allocation in stratified sampling methods. *Methodology and Computing in Applied Probability*, 12(3):335–360, 2010. doi: 10.1007/s11009-008-9108-0.

Granite Team, IBM. Granite-4.2-8B model card. Hugging Face model repository, 2026. URL https://huggingface.co/ibm-granite/granite-4.2-8b. Accessed Aug. 26, 2026.

Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995. doi: 10.1080/00401706.1995.10484303.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Sungee Hong, Zhengling Qi, and Raymond K. W. Wong. Distributional off-policy evaluation with bellman residual minimization. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4006–4014. PMLR, 2025.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. Active evaluation acquisition for efficient LLM benchmarking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35581–35602. PMLR, 2025.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Mistral AI. Mistral-Small-24B-Instruct-2501 model card. Hugging Face model repository, 2025. URL https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501. Accessed Aug. 31, 2026.

Subhojyoti Mukherjee, Josiah P. Hanna, and Robert D. Nowak. ReVar: Strengthening policy evaluation via reduced variance sampling. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1413–1422. PMLR, 2022.

Subhojyoti Mukherjee, Josiah P. Hanna, and Robert D. Nowak. SaVeR: Optimal data collection strategy for safe policy evaluation in tabular MDP. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36531–36576. PMLR, 2024.

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: 10.1111/j.2397-2335.1934.tb04184.x.

Phuc Nguyen, Deva Ramanan, and Charless Fowlkes. Active testing: An efficient and robust framework for estimating accuracy. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3759–3768. PMLR, 2018.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000. doi: 10.1080/01621459.2000.10473909.

Yang Peng and Liangyu Zhang. Online inference in distributional temporal-difference learning. *arXiv preprint arXiv:2608.14408*, 2026.

Yang Peng, Liangyu Zhang, and Zhihua Zhang. Statistical efficiency of distributional temporal difference learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24724–24761, 2024. doi: 10.52202/079017-0779.

Yang Peng, Kaicheng Jin, Liangyu Zhang, and Zhihua Zhang. A finite sample analysis of distributional temporal-difference learning with linear function approximation. *arXiv preprint arXiv:2502.14172*, 2025.

Qwen Team. Qwen3-32B model card. Hugging Face model repository, 2025a. URL https://huggingface.co/Qwen/Qwen3-32B. Accessed Aug. 31, 2026.

Qwen Team. Qwen3-4B-Instruct-2507 model card. Hugging Face model repository, 2025b. URL https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507. Accessed Aug. 21, 2026.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000. doi: 10.21314/JOR.2000.038.

Mark Rowland, Marc G. Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 29–37. PMLR, 2018.

Mark Rowland, Li Kevin Wenliang, Rémi Munos, Clare Lyle, Yunhao Tang, and Will Dabney. Near-minimax-optimal distributional reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 132774–132823, 2024. doi: 10.52202/079017-4221.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1468–1476, 2015.

Philip S. Thomas and Erik Learned-Miller. Concentration inequalities for conditional value at risk. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6225–6233. PMLR, 2019.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, pp. 163114–163129, 2025. doi: 10.52202/085713-4916.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 95266–95290, 2024. doi: 10.52202/079017-3018.

Runzhe Wu, Masatoshi Uehara, and Wen Sun. Distributional offline policy evaluation with predictive error guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37685–37712. PMLR, 2023.

Liangyu Zhang, Yang Peng, Jiadong Liang, Wenhao Yang, and Zhihua Zhang. Estimation and inference in distributional reinforcement learning. *The Annals of Statistics*, 53(5):1987–2011, 2025. doi: 10.1214/25-AOS2527.

Zhipu AI. GLM-4-32B-0414 model card. Hugging Face model repository, 2025. URL https://huggingface.co/zai-org/GLM-4-32B-0414. Accessed Aug. 31, 2026.

Yi Zhu, Jing Dong, and Henry Lam. Uncertainty quantification and exploration for reinforcement learning. *Operations Research*, 72(4):1689–1709, 2024. doi: 10.1287/opre.2023.2436.

SUPPLEMENT: TIS FOR CVAR POLICY EVALUATION

Appendices A–D develop the allocation theory from the stop-loss Bellman representation to fixed-design efficiency, learned oracle adaptation, and the anchored safeguard. The main results are proved as follows: Theorem 1 in Appendix C.1, Theorem 2 in Appendix C.2, Proposition 1 in Appendix C.4, Theorem 3 in Appendix D.3, the anchor's guarantees (Proposition 4 and Corollary 1) in Appendix D.4, and the pilot-payoff condition (Equation 11) in Appendix D.5. Appendix E closes the remaining links to the main paper: Appendix E.1 bounds categorical representation error; the experimental subsections give the protocols and evidence for **Q1–Q5**; and Appendix E.6 proves Proposition 2 and connects it to the tail–mean diagnostic. Section F positions these results against the closest foundations.

Proof dependencies. Lemma 1 justifies the stop-loss Bellman representation. Lemma 2 controls its empirical remainder and, together with the positive quantile margin, yields Theorem 1. That theorem identifies the influence used in the efficiency calculation of Theorem 2. Lemma 3 controls pilot estimates of the same influence; combined with Lemma 2, it yields Theorem 3. Proposition 4 then transfers the allocation control to the anchored design in Corollary 1.

A Notation and Assumptions **15**

B Allocation and Bellman Identities **16**

B.1 Minimizing the allocation variance 16

B.2 Projection and categorical CVaR identities 16

B.3 Exact empirical fixed-point expansion 17

C Fixed-Design Limits and Structural Interpretation **19**

C.1 Fixed-design CLT and normalized MSE 19

C.2 Semiparametric efficiency 20

C.3 Untied factorization for the allocation ablations 21

C.4 Controlled separation with fixed visitation, moments, and root law 22

D Learned Allocation **23**

D.1 Pilot-scale consistency and lower-tail control 23

D.2 Finite-pilot design stability 24

D.3 Oracle adaptation 25

D.4 Anchored design: variance safeguard and asymptotic MSE 26

D.5 When learning an allocation repays its pilot 28

E Approximation and Experimental Protocols **29**

E.1 Representation error 29

E.2 Experimental Evidence and Protocols 29

E.3 FinQA terminal-risk protocol and full results 37

E.4 Inventory disruption family 39

E.5 Blending controls 40

E.6 Rare failures and the tail–mean coincidence 40

E.7 FinQA with calculator faults 41

E.8 Longer review loops 42

E.9 Selecting the allocation from the pilot 42

E.10 Mechanism: what the pilot sees and what anchoring changes 43

F Additional Related Work **43**

A Notation and Assumptions

A query group identifies a sampled law; a Bellman row identifies one use of it. This section makes that distinction precise and connects the probability, measure, and shortfall notation used in the proofs.

Queryable-group experiment. $\mathcal{G}$ is a finite set of conditionally and independently queryable laws P_g, $G := |\mathcal{G}|$, and the objective is one fixed scalar functional of those laws. Section B.1 optimizes the variance once the groupwise influence scales are identified.

Categorical Bellman conditions. We evaluate one root (H, s_0) and let $\mathcal{B}$ index the required layer-state rows. A group may feed several rows through known nonnegative mixture coefficients summing to one within each Bellman row. A known component, if present, is represented by a degenerate query law with zero influence. Sub-probability row sums arise in the continuation matrix from termination at nonpositive shifted thresholds. The return law still retains all probability mass. In the stationary model, $g = (s, a)$, $W_g = (R, S') \sim P_{s,a}$, and the coefficient in row (h, s) is $\pi_h(a \mid s)$. The ordered grid $\mathcal{Z} = \{z_0, \dots, z_{K-1}\} \subset [0, H]$ contains both endpoints and has maximum gap Δ. The proofs keep H, $|\mathcal{S}|$, $|\mathcal{A}|$, and K fixed, assume bounded rewards, and impose the positive categorical quantile margin in Equation 21.

Untied special case. For independently queryable layer-specific laws, $g = b = (h, s)$ and P_b is the policy-mixture law obtained by drawing $A \sim \pi_h(\cdot \mid s)$ before the transition. One law then feeds one row. Structurally unreachable groups may be removed in either model only when the declared support and fixed policy certify their irrelevance. The retained rows must be closed under every possible continuation transition in the declared outcome spaces. These fixed outcome spaces also define the nonparametric product model in the efficiency theorem.

Equivalent probability and measure forms. The measure $\eta_h^*(s) = \sum_k p_{h,s,k}^* \delta_{z_k}$ is another notation for the grid probabilities in Section 2; hats denote the empirical version. The categorical target and estimator are

$$C_{\alpha,K} = \mathrm{CVaR}_\alpha\left(\eta_H^*(s_0)\right), \qquad \widehat{C}_N = \mathrm{CVaR}_\alpha\left(\widehat{\eta}_H(s_0)\right). \tag{12}$$

For the shift-and-project matrix in Equation 1, $Q(R)_{kj} = \ell_k(R + z_j)$, where $\ell_k(y)$ is the categorical projection weight on z_k. Column sums are one, including at clipped endpoints. Thus the vector equation is equivalent to

$$\eta_0^*(s) = \delta_0, \qquad \eta_h^*(s) = \sum_a \pi_h(a \mid s)\, \mathbb{E}_{(R,S')\sim P_{s,a}}\left[\Pi_C\left((f_R)_\# \eta_{h-1}^*(S')\right)\right], \quad f_R(z) = R + z. \tag{13}$$

Here $(f_R)_\#\nu$ is the law of $R + Z$ for $Z \sim \nu$. Replacing each expectation by the same group's empirical mean gives the estimator in Equation 1; the measure notation describes the same computation.

Stacked-coordinate convention and proof roadmap. We index a scalar shortfall (stop-loss) coordinate by $x = (h, s, q)$, write e_x for its standard basis vector, and write $U_{h,s} = U_h(s, \cdot) \in \mathbb{R}^K$ for the threshold block. A vector such as U stacks these coordinates; a row index $b = (h, s)$ selects one block. Sections B–C derive the influence and its sampling limits. Section D controls the extra error from learning the allocation.

For an outcome $W = (R, S')$ and stacked stop-loss vector U, define the sampled stop-loss Bellman target at row $b = (h, s)$ by

$$\left[T_b(W; U)\right](q) = \begin{cases} (q - R)_+, & h = 1, \\ \mathcal{I}_{\mathcal{Z}}\left[U_{h-1}(S', \cdot)\right](q - R), & h > 1, \end{cases} \tag{14}$$

where $\mathcal{I}_{\mathcal{Z}}$ is linear interpolation on the threshold grid, extended by zero for nonpositive arguments. For group g, let $\mathcal{T}_g(W_g; U)$ be its full stacked affine contribution, including all known mixture coefficients. In the shared-kernel model, for $g = (s, a)$ and $b = (h, s)$,

$$\left[\mathcal{T}_{s,a}(W_g; U)\right]_b = \pi_h(a \mid s) T_b(W_g; U),$$

and rows at states other than s are zero. Thus one W_g contributes jointly to every required layer row at state s; this is why its layer effects must be summed before their variance is computed. For a categorical continuation $Z \sim \eta_{h-1}^*(S')$, the stop-loss transform $v \mapsto \mathbb{E}[(v - Z)_+]$ is affine on every

grid interval $[z_j, z_{j+1}]$, and its values at the knots are exactly $U^*_{h-1}(S', z_j)$. Hence, for $0 < v \le H$, linear interpolation evaluates it exactly:

$$\mathcal{I}_{\mathcal{Z}}[U^*_{h-1}(S', \cdot)](v) = \mathbb{E}[(v - Z)_+].$$

Taking $v = q - R$ (with the stated zero extension when $q - R \le 0$) gives the shortfall of $R + Z$ at threshold q. Lemma 1 shows that categorical projection preserves this shortfall for every grid threshold q. Therefore

$$U^* = \sum_{g \in \mathcal{G}} \mathbb{E}_{P_g}\big[\mathcal{T}_g(W_g; U^*)\big] = \mathbf{b} + MU^*. \tag{15}$$

Set $A := I - M$. Because $q - R \le q \le H$, the finite-horizon target evaluates the interpolant only at or below the upper grid endpoint. The stated nonpositive extension is sufficient. In equation 15, $\mathbf{b}$ collects terms independent of continuation values, including the $h = 1$ rows. The matrix M collects the known mixture coefficients, transition expectations, and interpolation weights multiplying lower-layer coordinates. The matrix M lowers the layer and $M^H = 0$. Given n_g observations per group, replace each expectation by its empirical average. This is the stop-loss transform of the shared-kernel empirical categorical estimator. Denote the resulting affine map by $\widehat{\mathbf{b}} + \widehat{M}U$; again $\widehat{M}^H = 0$ for every dataset.

Explicit shared-kernel influence. Let $r_{h,s}$ be the adjoint block for layer–state row (h, s). For $g = (s, a)$, the full-vector formula in Equation 4 is

$$\phi_{s,a}(W) = -\frac{1}{\alpha} \sum_{h=1}^{H} \pi_h(a \mid s) r_{h,s}^{\top} \Big\{ T_h\big(W; U^*\big) - \mathbb{E}_{P_{s,a}} T_h\big(W; U^*\big) \Big\}. \tag{16}$$

The variance of this sum includes all cross-layer covariances induced by reusing the same observation.

B Allocation and Bellman Identities

Equation 6 follows from classical allocation once the scales are known. The identities below connect those scales to CVaR: projection preserves shortfalls, and Bellman propagation carries each group's error to the root.

B.1 Minimizing the allocation variance

Once the influence expansion gives $V(w) = \sum_g \sigma_g^2 / w_g$, classical Neyman allocation follows from Cauchy–Schwarz (Neyman, 1934):

$$\left(\sum_g \sigma_g \right)^2 = \left(\sum_g \frac{\sigma_g}{\sqrt{w_g}} \sqrt{w_g} \right)^2 \le \sum_g \frac{\sigma_g^2}{w_g}, \qquad \sum_g w_g = 1.$$

If all scales are positive, equality holds only at $w_g = \sigma_g / S_\sigma$, where $S_\sigma = \sum_g \sigma_g$. If $S_\sigma > 0$ but some scales vanish, this boundary design gives the infimum over positive designs: the mixtures $(1 - \lambda)\sigma / S_\sigma + \lambda \mathbf{1}/G$ approach it as $\lambda \downarrow 0$. If all scales vanish, every design has zero first-order variance. Sections B.3 and C.1 derive the influence expansion and its CLT and MSE limits for the categorical Bellman estimator.

B.2 Projection and categorical CVaR identities

Section 3 relies on projection preserving shortfalls and a stable quantile making CVaR locally affine. For return X with CDF F_X, lower-tail CVaR is

$$\mathrm{CVaR}_\alpha(X) = \frac{1}{\alpha} \int_0^\alpha F_X^{-1}(u)\, du, \quad \alpha \in (0, 1), \quad \text{with } F_X^{-1}(u) = \inf\{x : F_X(x) \ge u\}. \tag{17}$$

Lemma 1 (Projection identity). *For every $q \in \mathcal{Z}$ and every law ν supported on $[0, \infty)$,*

$$\int (q - z)_+ \, d\big(\Pi_C \nu\big)(z) = \int (q - z)_+ \, d\nu(z). \tag{18}$$

Proof. For $y \in [z_j, z_{j+1}]$, categorical projection (Bellemare et al., 2017; Rowland et al., 2018) sends δ_y to

$$\frac{z_{j+1} - y}{z_{j+1} - z_j}\delta_{z_j} + \frac{y - z_j}{z_{j+1} - z_j}\delta_{z_{j+1}}.$$

For a grid atom q, the map $z \mapsto (q - z)_+$ is affine on every grid cell. Its expectation is therefore preserved by the barycentric projection. If $y \geq H$, projection clips to $H \geq q$ and both the original and clipped payoffs are zero. Inputs below zero are excluded by the nonnegative-return model. Linearity proves the identity for every input law supported on $[0, \infty)$. □

For a categorical law p with CDF $F_k = \sum_{i \leq k} p_i$, set $F_{-1} = 0$, $k_\alpha = \min\{k : F_k \geq \alpha\}$, and $q_\alpha = z_{k_\alpha}$. The quantile-integral definition of lower-tail CVaR (Rockafellar & Uryasev, 2000) gives

$$\mathrm{CVaR}_\alpha(p) = \frac{1}{\alpha}\left\{\sum_{i<k_\alpha} p_i z_i + (\alpha - F_{k_\alpha - 1})z_{k_\alpha}\right\} \tag{19}$$

$$= z_{k_\alpha} - \frac{1}{\alpha}\sum_{i<k_\alpha} p_i(z_{k_\alpha} - z_i). \tag{20}$$

Thus CVaR is affine in a neighborhood where the VaR index is fixed.

For the population root law, let $F^*_k = \sum_{j \leq k} p^*_{H,s_0,j}$, $F^*_{-1} = 0$, and $k_\alpha = \min\{k : F^*_k \geq \alpha\}$. The positive quantile margin used in the main text is

$$m_\alpha = \min\left\{\alpha - F^*_{k_\alpha - 1},\ F^*_{k_\alpha} - \alpha\right\} > 0. \tag{21}$$

It places α strictly inside the quantile atom's cumulative-mass interval. Throughout the finite-horizon proof, $C_{\alpha,K} = q_\alpha - \alpha^{-1}U^*_H(s_0, q_\alpha)$ is the population categorical CVaR; $\widehat{C}_N$ uses the empirical fixed point and its empirical quantile atom.

Why a positive margin matters. For $H = 1$, take $Z \in \{0, 1\}$ with $p := \mathbb{P}(Z = 0)$. Then $\mathrm{CVaR}_\alpha(Z) = \max\{0, 1 - p/\alpha\}$. At $p = \alpha$ the margin vanishes, and for an empirical fraction $\widehat{p}$,

$$\sqrt{n}\,\widehat{C}_n = \max\{0, -\sqrt{n}(\widehat{p} - \alpha)/\alpha\} \Rightarrow \max\{0, -Z_0/\alpha\}, \qquad Z_0 \sim \mathcal{N}(0, \alpha(1 - \alpha)).$$

The limit has an atom at zero, so it is non-Gaussian. At $p > \alpha$ the margin is positive, CVaR is locally constant, and every influence is zero. The fixed-design theorem includes this degenerate limit. Oracle adaptation assumes $S_\sigma > 0$.

A shared-kernel example with nonzero covariance. Take one state, one action, $H = 2$, rewards $R \sim \mathrm{Bernoulli}(p)$, grid $\{0, 1, 2\}$, and $p = 1/2$. With $\alpha = 3/5$, the root VaR is 1 and the margin is $3/20$. Locally,

$$C_{\alpha,K} = 1 - \frac{(1 - p)^2}{\alpha}, \qquad \phi(R) = \frac{2(1 - p)}{\alpha}(R - p) = \frac{R - 1/2}{\alpha}.$$

Each of the two layer contributions is $(R - 1/2)/(2\alpha)$. Their sum has variance $1/(4\alpha^2) = 25/36$, whereas deleting their covariance gives $1/(8\alpha^2) = 25/72$. This demonstrates the variance error from treating a shared draw as two independent draws. There is only one group, so the example isolates covariance. The allocation itself is trivial.

B.3 Exact empirical fixed-point expansion

For Theorem 1, we separate leading sampling error from feedback due to estimating the continuation model. Define the centered contribution of one group observation

$$\Xi_g(W_g) = \mathcal{T}_g(W_g; U^*) - \mathbb{E}_{P_g}\mathcal{T}_g(W_g; U^*), \qquad \mathbb{E}_{P_g}\Xi_g = 0, \tag{22}$$

and the combined empirical Bellman error

$$\widehat{\xi} = \sum_{g \in \mathcal{G}} \frac{1}{n_g}\sum_{i=1}^{n_g} \Xi_g(W_{g,i}). \tag{23}$$

By the affine form of $\mathcal{T}_g$, this same perturbation can be written as

$$\widehat{\xi} = \Big(\widehat{\mathbf{b}} - \mathbf{b}\Big) + \Big(\widehat{M} - M\Big)U^*.$$

Thus $\widehat{\xi}$ is the empirical Bellman error evaluated at the population fixed point. The resolvent below converts it into fixed-point estimation error. Subtracting the population and empirical fixed-point equations gives the exact identity

$$\widehat{U} - U^* = \Big(I - \widehat{M}\Big)^{-1}\widehat{\xi}. \tag{24}$$

Indeed,

$$\begin{aligned}\Big(I - \widehat{M}\Big)\Big(\widehat{U} - U^*\Big) &= \widehat{\mathbf{b}} + \widehat{M}U^* - U^* \\ &= \Big(\widehat{\mathbf{b}} - \mathbf{b}\Big) + \Big(\widehat{M} - M\Big)U^* = \widehat{\xi}.\end{aligned}$$

Both inverses are finite sums:

$$\Big(I - \widehat{M}\Big)^{-1} = \sum_{t=0}^{H-1} \widehat{M}^t, \qquad \Big(I - M\Big)^{-1} = \sum_{t=0}^{H-1} M^t.$$

Every row of M and $\widehat{M}$ is sub-probability, so their induced infinity norms are at most one and both inverse norms are at most H.

The resolvent identity gives

$$\widehat{U} - U^* = A^{-1}\widehat{\xi} + R_N, \tag{25}$$

$$R_N = A^{-1}\Big(\widehat{M} - M\Big)\Big(I - \widehat{M}\Big)^{-1}\widehat{\xi}. \tag{26}$$

In the finite-horizon bounds below, $\|\cdot\|$ denotes the vector infinity norm or its induced matrix infinity norm. With $n_{\min} = \min_g n_g$, finite dimension and bounded observations imply

$$\left\|\widehat{M} - M\right\| = O_p\left(n_{\min}^{-1/2}\right), \quad \left\|\widehat{\xi}\right\| = O_p\left(n_{\min}^{-1/2}\right), \quad \|R_N\| = O_p\left(n_{\min}^{-1}\right). \tag{27}$$

Lemma 2 (Uniform moment control)**.** *For every fixed integer $p \geq 2$, there is a finite constant C_p, depending only on the fixed model dimensions, grid, horizon, and p, such that*

$$\begin{aligned}\mathbb{E}\left\|\widehat{M} - M\right\|^p + \mathbb{E}\left\|\widehat{\xi}\right\|^p &\leq C_p n_{\min}^{-p/2}, \\ \mathbb{E}\,\|R_N\|^p &\leq C_p n_{\min}^{-p}.\end{aligned} \tag{28}$$

In particular, under a stable allocation, $N\mathbb{E}\|R_N\|^2 \to 0$.

Proof. **Idea.** Both empirical coefficient error and Bellman error are bounded sample averages, hence of order $n_{\min}^{-1/2}$. The fixed-point remainder is their product, so it is one order smaller. The deterministic resolvent bound prevents the recursion from amplifying these rates.

Each entry of $\widehat{M} - M$ and $\widehat{\xi}$ is a finite sum, over groups, of centered averages of bounded random variables. For a centered average $\bar{X}_g$ of n_g independent bounded variables, Hoeffding's inequality (Hoeffding, 1963) gives $\mathbb{P}(|\bar{X}_g| > t) \leq 2e^{-cn_g t^2}$. Integrating this tail via $\mathbb{E}|\bar{X}_g|^p = \int_0^\infty pt^{p-1}\mathbb{P}(|\bar{X}_g| > t)\,dt$ yields $\mathbb{E}|\bar{X}_g|^p \leq C_p n_g^{-p/2}$. The number of groups and matrix entries is fixed, so norm equivalence and a finite-sum inequality give the first line of equation 28. Nilpotence and the sub-probability row structure give the deterministic bounds $\|A^{-1}\| \leq H$ and $\|(I - \widehat{M})^{-1}\| \leq H$. Hence

$$\|R_N\| \leq H^2\left\|\widehat{M} - M\right\|\left\|\widehat{\xi}\right\|.$$

Cauchy–Schwarz with the preceding $2p$-moment bounds proves the second line. If $n_{\min} \asymp N$, it yields $N\mathbb{E}\|R_N\|^2 = O(N^{-1})$. □

C Fixed-Design Limits and Structural Interpretation

Theorems 1 and 2 turn the influence calculation into an attainable accuracy benchmark for fixed query shares. The untied factorization explains the allocation ablations; Proposition 1 then shows why visitation and reward moments cannot determine the best tail allocation.

C.1 Fixed-design CLT and normalized MSE

Proof of Theorem 1. **Idea.** There are three issues to separate: linearize the empirical Bellman fixed point, show the empirical VaR atom stays on the same categorical cell, and then transfer the group-wise CLT and second moment through that locally affine CVaR readout.

Step 1: asymptotic linearity on the correct VaR cell. Let E_N be the event that the empirical and population categorical VaR indices agree, set $x_0 = (H, s_0, q_\alpha)$ and $r^\top = e_{x_0}^\top A^{-1}$, and define

$$\phi_g(W_g) := -\alpha^{-1} r^\top \Xi_g(W_g), \qquad L_N := \sum_g \frac{1}{n_g} \sum_{i=1}^{n_g} \phi_g(W_{g,i}). \tag{29}$$

Then $\mathbb{E}_{P_g}\phi_g = 0$ and $\mathbb{E}_{P_g}\phi_g^2 = \sigma_g^2$. On E_N, equation 20 and equation 25 give

$$\widehat{C}_N - C_{\alpha,K} = L_N + \widetilde{R}_N, \qquad \widetilde{R}_N := -\alpha^{-1} e_{x_0}^\top R_N. \tag{30}$$

Under a stable allocation, $n_{\min} := \min_g n_g \asymp N$, so Lemma 2 gives $\sqrt{N}\widetilde{R}_N = o_p(1)$. For each fixed group,

$$\frac{\sqrt{N}}{n_g} \sum_{i=1}^{n_g} \phi_g(W_{g,i}) = \sqrt{\frac{N}{n_g}} \frac{1}{\sqrt{n_g}} \sum_{i=1}^{n_g} \phi_g(W_{g,i}) \Rightarrow \mathcal{N}\left(0, \frac{\sigma_g^2}{w_g}\right),$$

by the ordinary CLT and $n_g/N \to w_g > 0$. The groups are independent and their number is fixed, so the vector of group terms converges jointly to independent Gaussian limits. Summing the coordinates gives

$$\sqrt{N} L_N \Rightarrow \mathcal{N}(0, V(w)), \qquad V(w) := \sum_g \frac{\sigma_g^2}{w_g}. \tag{31}$$

Step 2: the VaR cell is correct with exponentially high probability. Each entry of $\widehat{\mathbf{b}} - \mathbf{b}$ and $\widehat{M} - M$ is a finite sum of averages of bounded variables. Hoeffding's inequality (Hoeffding, 1963) and a union bound therefore give constants $c_1, c_2 > 0$ such that

$$\mathbb{P}\left(\|\widehat{\mathbf{b}} - \mathbf{b}\|_\infty + \|\widehat{M} - M\|_\infty > t\right) \le c_1 e^{-c_2 n_{\min} t^2}, \qquad t > 0. \tag{32}$$

For a categorical law with stop-loss vector U and CDF F,

$$F_{j-1} = \frac{U(z_j) - U(z_{j-1})}{z_j - z_{j-1}} \quad (j = 1, \dots, K-1), \qquad F_{K-1} = 1.$$

Because every shortfall coordinate satisfies $0 \le U_h^*(s,q) \le q \le H$, we have $\|U^*\|_\infty \le H$. Thus, with $\delta_{\min} := \min_{j<K-1}(z_{j+1} - z_j) > 0$, the exact fixed-point identity and $\|(I - \widehat{M})^{-1}\|_\infty \le H$ imply

$$\begin{aligned} \|\widehat{U} - U^*\|_\infty &\le H\{\|\widehat{\mathbf{b}} - \mathbf{b}\|_\infty + H\|\widehat{M} - M\|_\infty\}, \\ \|\widehat{F} - F^*\|_\infty &\le \frac{2H}{\delta_{\min}}\{\|\widehat{\mathbf{b}} - \mathbf{b}\|_\infty + H\|\widehat{M} - M\|_\infty\}. \end{aligned} \tag{33}$$

Combining this bound with equation 32 gives

$$\mathbb{P}(E_N^c) \le \mathbb{P}(\|\widehat{F} - F^*\|_\infty \ge m_\alpha) \le c_3 e^{-c_4 n_{\min} m_\alpha^2}. \tag{34}$$

Indeed, an error smaller than m_α preserves $\widehat{F}_{k_\alpha - 1} < \alpha < \widehat{F}_{k_\alpha}$; at the endpoints use $\widehat{F}_{-1} = 0$ and $\widehat{F}_{K-1} = 1$.

Step 3: normalized MSE and transfer off the good event. For the second-moment claim, centering and independence give

$$N\mathbb{E}L_N^2 = N\sum_g \frac{\sigma_g^2}{n_g} \longrightarrow V(w). \tag{35}$$

Lemma 2 also gives $N\mathbb{E}\widetilde{R}_N^2 = O(N^{-1})$ and, by Cauchy–Schwarz, $N|\mathbb{E}[L_N\widetilde{R}_N]| = o(1)$. Hence $N\mathbb{E}(L_N + \widetilde{R}_N)^2 \to V(w)$. All three variables $\widehat{C}_N - C_{\alpha,K}$, L_N, and $\widetilde{R}_N$ are uniformly bounded; for the last, use equation 26 and the deterministic resolvent bounds. Therefore

$$N\mathbb{E}\Big[\{(\widehat{C}_N - C_{\alpha,K})^2 + (L_N + \widetilde{R}_N)^2\}\mathbf{1}_{E_N^c}\Big] \le CN\mathbb{P}(E_N^c) \longrightarrow 0. \tag{36}$$

Equation 30 holds on E_N, so the last two displays prove the normalized-MSE limit. Since $\mathbb{P}(E_N^c) \to 0$ and $\sqrt{N}\widetilde{R}_N = o_p(1)$, they also transfer the CLT for L_N to $\widehat{C}_N$. □

C.2 Semiparametric efficiency

Proof of Theorem 2. **Idea.** Differentiate the target along arbitrary groupwise score directions. The resulting pathwise derivative is represented by ϕ_g in each group. Under sampling fraction w_g, the product-experiment canonical gradient is therefore ϕ_g/w_g, whose squared norm is exactly $V(w)$. The asymptotic-linear expansion from Theorem 1, together with Le Cam's third lemma, then shows that the plug-in estimator is regular under local alternatives and attains this bound.

Step 1: pathwise derivative. Fix the population collection $P = (P_g)_{g\in\mathcal{G}}$ and let $\mathcal{P} = \bigotimes_{g\in\mathcal{G}} \mathcal{P}_g$, where each $\mathcal{P}_g$ is the nonparametric model on group g's fixed bounded outcome space. Write $L_0^2(P_g)$ for the square-integrable, mean-zero functions under P_g. For a bounded $s_g \in L_0^2(P_g)$ and sufficiently small $|t|$, $dP_{g,t} = (1 + ts_g)dP_g$ is a valid submodel with score s_g; bounded mean-zero scores are dense in $L_0^2(P_g)$ (van der Vaart, 1998, Chapters 7 and 25). It therefore suffices to derive the gradient first for bounded scores; because all Bellman contributions and the resulting ϕ_g are bounded, the derivative extends continuously to the $L_0^2(P_g)$ closure.

Write the affine group contribution as $\mathcal{T}_g(W;U) = a_g(W) + B_g(W)U$. Along simultaneous differentiable-in-quadratic-mean paths with scores $s_g \in L_0^2(P_g)$, differentiating equation 15 at $t = 0$ gives

$$\dot{U}_s = \sum_g \mathbb{E}_{P_g}[\mathcal{T}_g(W_g;U^*)s_g(W_g)] + M\dot{U}_s.$$

Because $\mathbb{E}_{P_g} s_g = 0$, the expectation in brackets equals $\mathbb{E}_{P_g}[\Xi_g(W_g)s_g(W_g)]$. Hence

$$A\dot{U}_s = v_s, \qquad v_s := \sum_g \mathbb{E}_{P_g}[\Xi_g(W_g)s_g(W_g)]. \tag{37}$$

The positive margin fixes the root VaR index on a neighborhood of the population law, so the CVaR readout is locally affine. Therefore

$$\begin{aligned}\dot{C}_s &= -\frac{1}{\alpha}e_{x_0}^\top \dot{U}_s = -\frac{1}{\alpha}r^\top v_s \\ &= \sum_g \mathbb{E}_{P_g}[\phi_g(W_g)s_g(W_g)].\end{aligned} \tag{38}$$

Thus $C_{\alpha,K}$ is pathwise differentiable at the population law with groupwise derivative representers ϕ_g.

Step 2: canonical gradient under the sampling design. For the deterministic allocation, applying local asymptotic normality groupwise to $P_{g,t/\sqrt{N}}^{\otimes n_g}$ and summing the log-likelihood ratios gives a product LAN experiment (van der Vaart, 1998, Theorem 7.2) with tangent inner product

$$\langle s, \widetilde{s}\rangle_w := \sum_g w_g \mathbb{E}_{P_g}[s_g\widetilde{s}_g].$$

By equation 38, its Riesz representer is $\psi_{w,g} = \phi_g / w_g$, because $\langle \psi_w, s \rangle_w = \sum_g \mathbb{E}_{P_g}[\phi_g s_g] = \dot{C}_s$. Its squared norm is

$$\|\psi_w\|_w^2 = \sum_g w_g \mathbb{E}_{P_g}\left[\left(\frac{\phi_g}{w_g}\right)^2\right] = \sum_g \frac{\sigma_g^2}{w_g} = V(w). \tag{39}$$

Step 3: regularity and attainment under local alternatives. Theorem 1 gives the baseline asymptotic-linear representation

$$\sqrt{N}\{\widehat{C}_N - C_{\alpha,K}(P)\} = \sum_g \frac{\sqrt{N}}{n_g} \sum_{i=1}^{n_g} \phi_g(W_{g,i}) + o_P(1) = \frac{1}{\sqrt{N}} \sum_g \sum_{i=1}^{n_g} \psi_{w,g}(W_{g,i}) + o_P(1).$$

For the last relation, write $N/n_g = w_g^{-1} + o(1)$. Since, for each fixed group, $N^{-1/2} \sum_{i=1}^{n_g} \phi_g(W_{g,i}) = O_P(1)$, replacing N/n_g by $1/w_g$ changes the finite sum by only $o_P(1)$. Fix a DQM direction $s = (s_g)$ and a scalar t. Write $P_{t/\sqrt{N}} := (P_{g,t/\sqrt{N}})_{g \in \mathcal{G}}$ for the local collection of group laws and $P_{N,t} := \bigotimes_g P_{g,t/\sqrt{N}}^{\otimes n_g}$ for the corresponding sample law. LAN implies that $P_{N,t}$ is contiguous to the baseline sample law, so the displayed $o_P(1)$ term is also $o_{P_{N,t}}(1)$.

By Step 1,

$$\sqrt{N}\{C_{\alpha,K}(P_{t/\sqrt{N}}) - C_{\alpha,K}(P)\} \longrightarrow t\dot{C}_s.$$

Under the baseline law, the joint CLT for the influence sum and the LAN central sequence has covariance $\langle \psi_w, s \rangle_w = \dot{C}_s$. Le Cam's third lemma (van der Vaart, 1998, Chapter 6) therefore gives, under $P_{N,t}$,

$$\frac{1}{\sqrt{N}} \sum_g \sum_{i=1}^{n_g} \psi_{w,g}(W_{g,i}) \Rightarrow \mathcal{N}(t\dot{C}_s, V(w)).$$

Subtracting the local target shift yields

$$\sqrt{N}\{\widehat{C}_N - C_{\alpha,K}(P_{t/\sqrt{N}})\} \Rightarrow \mathcal{N}(0, V(w)) \qquad \text{under } P_{N,t}. \tag{40}$$

This limit is the same for every fixed t and DQM direction, which is the required regularity. Hence the plug-in estimator attains the canonical-gradient variance $V(w)$.

Step 4: lower bound. The product tangent space is linear and hence a convex cone. The convolution theorem therefore implies that the limit law of any regular estimator is the convolution of $\mathcal{N}(0, V(w))$ with an independent remainder; in particular, whenever its second moment is finite, its asymptotic variance is at least $V(w)$ (van der Vaart, 1998, Theorem 25.20). The local asymptotic minimax theorem gives the corresponding lower bound $V(w)$ for local squared-error risk (van der Vaart, 1998, Theorem 25.21). The plug-in estimator has the Gaussian limit in equation 40, so it attains both bounds. Thus it is semiparametrically efficient for the fixed design. □

Section B.1 minimizes this bound, including zero-scale groups.

C.3 Untied factorization for the allocation ablations

The ablations in Figure 3 separate propagation from local variability. When one independent law feeds each layer–state row $b = (h, s)$, its centered innovation is $\zeta_b(W) = T_b(W; U^*) - U_b^*$. Thus

$$\phi_b(W) = -\alpha^{-1} r_b^\top \zeta_b(W).$$

The adjoint is nonnegative because $r^\top = e_{x_0}^\top \sum_{j=0}^{H-1} M^j$ and $M \geq 0$. Set $d_b = \mathbf{1}^\top r_b$. If $d_b > 0$, normalize its threshold weights as $\rho_b = r_b / d_b$ and define

$$\tau_b^2 = \alpha^{-2} \operatorname{Var}_{P_b}(\rho_b^\top \zeta_b(W)), \qquad \sigma_b = d_b \tau_b.$$

If $d_b = 0$, then $r_b = 0$ and $\sigma_b = 0$; set $\tau_b = 0$. This is the factorization used by the reachability-only and local-scale-only ablations. The quantity d_b is adjoint mass: it propagates visitation together with the remaining shortfall threshold, so it need not equal ordinary state occupancy. The factor τ_b measures variability after averaging over those threshold weights. The oracle uses their product.

C.4 Controlled separation with fixed visitation, moments, and root law

Proposition 1 isolates tail information missing from visitation and reward moments while preserving the root law. Section E.2.2 tests whether a charged pilot learns this population signal.

Proof of Proposition 1. Consider one initial state, $H = 1$, $G \geq 2$ actions with probabilities $1/G$, and a fixed terminal next state. Each action's reward law is independently queryable. Set $\alpha = 1/10$ and use the grid $\{0, 1/3, 5/9, 2/3, 1\}$, which contains every reward below and makes projection exact. Define two reward laws:

$$P^{\mathrm{A}}: \quad \mathbb{P}(R=0) = \frac{1}{10}, \quad \mathbb{P}(R=5/9) = \frac{9}{10}; \qquad P^{\mathrm{B}}: \quad \mathbb{P}(R=1/3) = \mathbb{P}(R=2/3) = \frac{1}{2}. \tag{41}$$

Both have $\mathbb{E}R = 1/2$ and $\mathbb{E}R^2 = 5/18$, hence $\mathrm{Var}(R) = 1/36$. These are properties of the construction; estimation still uses the unrestricted group-law model. Assign P^{A} to group 1 and P^{B} to every other group; denote these separated laws by P_g^{sep}.

The root law is $\overline{P} = G^{-1}P^{\mathrm{A}} + (1 - G^{-1})P^{\mathrm{B}}$. Its mass strictly below $1/3$ is $1/(10G)$ and its cumulative mass at $1/3$ is $1/2 - 2/(5G)$. Thus, for every $G \geq 2$,

$$q_\alpha = \frac{1}{3}, \qquad m_\alpha = \frac{1}{10}\left(1 - \frac{1}{G}\right) > 0, \qquad C_{\alpha,K} = \frac{G-1}{3G}. \tag{42}$$

Let $L(R) = (1/3 - R)_+$. The root is the known uniform mixture of the group laws, so its group influence is

$$\phi_g(R) = -\frac{1}{G\alpha}\{L(R) - \mathbb{E}_{P_g} L(R)\}. \tag{43}$$

Under P^{A}, L is $1/3$ with probability $1/10$ and zero otherwise, giving $\mathrm{Var}(L) = 1/100$. Under P^{B}, L is identically zero. Consequently $\sigma_1 = 1/G$ and $\sigma_g = 0$ for $g > 1$. The occupancy design is uniform. The group influence for estimating the mean is $(R - 1/2)/G$, with standard deviation $1/(6G)$ in every group, so the mean-optimal design is also uniform. Evaluating CVaR variance under either design gives $V = G\sum_g \sigma_g^2 = 1/G$, whereas $V^* = (\sum_g \sigma_g)^2 = 1/G^2$. The tail oracle is a boundary design; its value is approached by positive designs as the floor vanishes.

For the fixed-root-law family, set

$$P_g(t) = (1-t)\overline{P} + tP_g^{\mathrm{sep}}, \qquad 0 \leq t \leq 1. \tag{44}$$

Every component being mixed has the same first two reward moments, so each $P_g(t)$ retains mean $1/2$ and variance $1/36$. Moreover, $G^{-1}\sum_g P_g(t) = \overline{P}$ for every t: the full root return law, CVaR, and quantile margin in Equation 42 stay fixed. Writing $\theta_g(t) = P_g(t)\{R = 0\}$ gives

$$\theta_1(t) = \frac{1 + (G-1)t}{10G}, \qquad \theta_g(t) = \frac{1-t}{10G}\ (g > 1), \qquad \sigma_g(t) = \frac{10}{3G}\sqrt{\theta_g(t)(1 - \theta_g(t))}. \tag{45}$$

For uniform occupancy, the variance ratio is explicitly

$$\frac{V_{\mathrm{occupancy}}(t)}{V^*(t)} = \frac{G\sum_g \sigma_g(t)^2}{\left(\sum_g \sigma_g(t)\right)^2}.$$

Cauchy–Schwarz gives the lower bound 1, while nonnegativity gives $\sum_g \sigma_g(t)^2 \leq (\sum_g \sigma_g(t))^2$ and hence the upper bound G. At $t = 0$ all scales are equal and positive, so the ratio is one. At $t = 1$ it is G, as above. For all intermediate t the ratio is continuous. It therefore attains every value between the two endpoints. Every scale is positive when $t < 1$, so ratios arbitrarily close to G also occur with all groups influential. This proves the family without a change in visitation, conditional moments, or root return law. □

Rollouts and the ideal occupancy anchor. In this one-step example one rollout and one conditional observation each cost one transition query. For complete rollouts, the shortfall indicator has probability $1/(10G)$ under the fixed root law. The empirical CVaR therefore has leading variance

$$V_{\mathrm{rollout}} = \frac{(1/3)^2}{\alpha^2}\frac{1}{10G}\left(1 - \frac{1}{10G}\right) = \frac{10G - 1}{9G^2}, \tag{46}$$

which does not change with t. At $t = 1$, the equal mixture of the population tail oracle and uniform occupancy gives group 1 weight $(G+1)/(2G)$ and hence $V_{\text{ideal anchor}} = 2/[G(G+1)]$. For $G = 10$, the variance constants for occupancy, rollouts, the tail oracle, and this ideal anchor are respectively .1, .11, .01, and $1/55$. These are leading-variance constants at the separated endpoint, before pilot cost, exploration, and rounding. For intermediate t, use Equation 45; the occupancy variance itself changes with t even though the root law is fixed.

From population allocation to pilot learning. At $t = 1$, m independent pilot observations from group 1 miss its zero reward with probability $(9/10)^m$. This is a missed-outcome probability, not the probability of TIS failure or of a wrong pilot quantile. Section E.2.2 reports learned performance with all queries charged and the prescribed uniform fallback.

D Learned Allocation

Algorithm 1 Tail-Influence Sampling (TIS); pilot cost is included in N

Require: query groups $\mathcal{G}$, budget N, grid $\mathcal{Z}$, tail level α; pilot size $m_N \geq 2$, floor $0 < \lambda_N < 1$, $N - Gm_N \geq 2G$

1: Draw m_N independent pilot samples per group; fit empirical conditional laws.
2: Compute the pilot root quantile, shortfalls, and adjoint via Equations 1 and 3.
3: Score each pilot draw and compute $\widehat{\sigma}_g$ by Equation 7.
4: Form $\widehat{w}$ by Equation 8; if all scales vanish, use $\widehat{w}_g = 1/G$.
5: Set $n_g = 2 + \mathrm{LRM}_g(N - Gm_N - 2G, \widehat{w})$.
6: Draw n_g fresh main samples per group; discard the pilot.
7: Evaluate Equation 1 with the main empirical laws and return the root CVaR $\widehat{C}_N^{\mathrm{TIS}}$.

Algorithm 1 learns allocation scales using a pilot, as in adaptive stratified sampling (Étoré & Jourdain, 2010; Carpentier et al., 2015), but also estimates the Bellman model and quantile. We control these errors to prove oracle adaptation, then establish the anchor's separate safeguard. Only fresh main samples form the final estimate; $G = |\mathcal{G}|$.

D.1 Pilot-scale consistency and lower-tail control

We use the centered form of Equation 7. For $m \geq 2$ pilot draws per group, set $\widehat{x}_0 = (H, s_0, \widehat{q}_\alpha^{(0)})$, the root coordinate at the pilot quantile. Define

$$\begin{aligned}
\widehat{r}^{(0)\top} &:= e_{\widehat{x}_0}^\top (I - \widehat{M}^{(0)})^{-1}, \\
\widehat{\Xi}_{g,i}^{(0)} &= \mathcal{T}_g(W_{g,i}; \widehat{U}^{(0)}) - \frac{1}{m}\sum_{j=1}^{m} \mathcal{T}_g(W_{g,j}; \widehat{U}^{(0)}), \\
\widehat{\phi}_{g,i}^{(0)} &= -\alpha^{-1} \widehat{r}^{(0)\top} \widehat{\Xi}_{g,i}^{(0)}, \\
\widehat{\sigma}_g^2 &= \frac{1}{m}\sum_{i=1}^{m} (\widehat{\phi}_{g,i}^{(0)})^2.
\end{aligned} \tag{47}$$

The four lines give, respectively, weights that propagate local changes to the root shortfall, each draw's deviation from its group's mean Bellman contribution, its estimated CVaR influence, and the empirical variance of these influences. Linearity gives $\widehat{\phi}_{g,i}^{(0)} = \widehat{d}_{g,i} - \bar{d}_g$, recovering Equation 7. Take $m = m_N$ for Theorem 3. Shared stage effects are summed before taking the variance, preserving their covariance.

Lemma 3 (**Pilot-scale control**). *Under the fixed-dimensional bounded Bellman model and positive quantile margin, there is a deterministic $B_\sigma < \infty$ such that $0 \leq \widehat{\sigma}_g \leq B_\sigma$ for every group and pilot dataset. With each pilot formed from the first m observations of an i.i.d. stream in each group,*

$$\widehat{\sigma}_g \longrightarrow \sigma_g \quad \text{a.s. as } m \to \infty. \tag{48}$$

For each group with $\sigma_g > 0$, constants $c_g, C_g > 0$ exist such that

$$\mathbb{P}(\widehat{\sigma}_g < \sigma_g/2) \leq C_g e^{-c_g m}. \tag{49}$$

Proof. **Idea.** In fixed dimension, the pilot scale is a continuous function of finitely many bounded empirical moments as long as the VaR cell is correct. Laws of large numbers give consistency, while Hoeffding bounds plus the positive margin control the rare event that a genuinely positive scale is badly underestimated.

Step 1: finite empirical-moment representation. Write the affine group contribution as $\mathcal{T}_g(W;U) = a_g(W) + B_g(W)U$. Let $Y_g(W)$ collect the finitely many entries of $a_g(W)$ and $B_g(W)$ and their pairwise products. The empirical first moments of a_g and B_g determine $\widehat{\mathbf{b}}^{(0)}$ and $\widehat{M}^{(0)}$, hence $\widehat{U}^{(0)}$ and $\widehat{r}^{(0)}$. Expanding the empirical variance in equation 47 then introduces only empirical averages of pairwise products of entries of a_g and B_g; no higher empirical moments are needed. These features are bounded, and all quantities in equation 47 are therefore functions of their groupwise empirical means $\widehat{\mu}$; let μ denote the corresponding vector of population means. On the correct VaR cell,

$$\widehat{U}^{(0)} = \sum_{j=0}^{H-1} (\widehat{M}^{(0)})^j \widehat{\mathbf{b}}^{(0)}, \qquad \widehat{r}^{(0)\top} = e_{\widehat{x}_0}^\top \sum_{j=0}^{H-1} (\widehat{M}^{(0)})^j,$$

so $\widehat{\sigma}_g^2 = F_g(\widehat{\mu})$ for a polynomial F_g with $F_g(\mu) = \sigma_g^2$.

Step 2: consistency. The strong law gives $\widehat{\mu} \to \mu$ almost surely. The positive margin and the CDF bound used in equation 34 make the pilot VaR cell eventually correct almost surely. Continuity of F_g proves equation 48. Bounded features, stop-loss coordinates, and the deterministic resolvent bound also give the uniform constant B_σ.

Step 3: lower-tail protection for active groups. If $\sigma_g > 0$, F_g is Lipschitz on a compact neighborhood of μ. Choose that neighborhood so $|F_g(\widehat{\mu}) - \sigma_g^2| \le 3\sigma_g^2/4$. Hoeffding's inequality (Hoeffding, 1963) and a finite union bound show that leaving this neighborhood has probability at most Ce^{-cm}. The same bound holds for a wrong pilot VaR cell by equation 34. Off these two events, $\widehat{\sigma}_g^2 \ge \sigma_g^2/4$, proving equation 49. □

The constants in Equation 49 depend on the fixed population model, including its positive influence scales and quantile margin. The bound supports the asymptotic MSE proof; it does not by itself give a pilot size computable from the data or a bound on the final estimator's finite-sample MSE.

Matrix-free influence computation. The displayed matrices define the linear operator. A matrix-free implementation avoids forming A^{-1} or the $D \times D$ covariance matrices. Compute the stop losses in increasing layer order, propagate the root adjoint in decreasing layer order, and accumulate the scalar $r^\top \mathcal{T}_g(W;U)$ for each sample before estimating its variance. In the shared state–action model, a straightforward sample-based implementation uses at most $O(HKN \log K + H|\mathcal{S}|K)$ arithmetic operations for these passes with binary search on a nonuniform grid, and $O(H|\mathcal{S}|K + N + G)$ storage. Interpolation indices can be reused. These are upper bounds for the described construction. Measured runtimes depend on the actual grid and implementation and are reported with the experimental artifacts.

D.2 Finite-pilot design stability

Near an interior oracle, small score errors have a quadratic variance cost. Severe underestimation requires the asymptotic controls in Section D.3.

Proposition 3 (Local design stability)**.** *If every retained $\sigma_g > 0$ and $\epsilon = \max_g |\widehat{\sigma}_g - \sigma_g|$, then for sufficiently small ϵ and floor λ,*

$$0 \le V(\widehat{w}) - V^* \le C(\epsilon^2 + \lambda^2) \tag{50}$$

for a finite problem-dependent constant C.

Proof. **Idea.** At an interior oracle, allocation error has a quadratic variance cost. Set $S_\sigma = \sum_g \sigma_g$, $p_g = \sigma_g/S_\sigma$, and $p_{\min} = \min_g p_g > 0$. For every positive design w,

$$V(w) - V^* = S_\sigma^2 \sum_g \frac{(p_g - w_g)^2}{w_g}. \tag{51}$$

To verify the identity, expand the square and use $\sum_g p_g = \sum_g w_g = 1$. If $G\epsilon \le S_\sigma/2$, normalizing the estimated scales gives

$$\left\| \frac{\widehat{\sigma}}{\sum_g \widehat{\sigma}_g} - p \right\|_\infty \le \frac{2(G+1)\epsilon}{S_\sigma}.$$

Adding the uniform floor therefore gives $\|\widehat{w} - p\|_\infty \le 2(G+1)\epsilon/S_\sigma + \lambda$. For sufficiently small ϵ, λ, every denominator $\widehat{w}_g$ in Equation 51 is at least $p_{\min}/2$. Substitution and $(a+b)^2 \le 2a^2 + 2b^2$ prove the claim. The lower bound on the weights is local; it gives no protection on a pilot event with severe scale underestimation. □

D.3 Oracle adaptation

Proof of Theorem 3. **Idea.** The pilot must make the positive-scale weights converge to the oracle design and make severe underestimation sufficiently rare for second moments. The exploration floor separately guarantees enough main samples in every group to control the nonlinear fixed-point remainder and the VaR cell. Conditional on the pilot, the remaining problem is a deterministic triangular-array CLT.

Step 1: learned weights, rounding, and a minimum main count. Write $S_\sigma := \sum_g \sigma_g > 0$ and $w_g^* := \sigma_g/S_\sigma$. Let $\widehat{C}_N^{\mathrm{TIS}}$ denote the main-sample estimator produced by Algorithm 1. Define

$$\widetilde{p}_g := \begin{cases} \widehat{\sigma}_g / \sum_j \widehat{\sigma}_j, & \sum_j \widehat{\sigma}_j > 0, \\ 1/G, & \text{otherwise}, \end{cases} \qquad \widehat{w}_g := (1-\lambda_N)\widetilde{p}_g + \lambda_N/G.$$

The second branch also gives $\widehat{w}_g = 1/G$. Pilot consistency and $S_\sigma > 0$ imply that its probability tends to zero. For pilots redrawn at each budget, the convergence needed below is in probability:

$$\widehat{w}_g \xrightarrow{p} w_g^* \quad \text{for every group, including } w_g^* = 0. \tag{52}$$

Set $J_N := N - Gm_N - 2G$ and $n_g := 2 + \mathrm{LRM}_g(J_N, \widehat{w})$. Largest-remainder rounding starts from $\lfloor J_N \widehat{w}_g \rfloor$ and assigns the leftover calls to the largest fractional remainders, with ties broken in a fixed group order. It satisfies

$$\sum_g \mathrm{LRM}_g(J_N, \widehat{w}) = J_N, \qquad |\mathrm{LRM}_g(J_N, \widehat{w}) - J_N \widehat{w}_g| < 1.$$

Thus the pilot and main counts sum to N. Since $J_N/N \to 1$, uniformly in g,

$$\frac{n_g}{N} - \widehat{w}_g = \left(\frac{J_N}{N} - 1\right)\widehat{w}_g + O(N^{-1}) = o(1). \tag{53}$$

Moreover, $\widehat{w}_g \ge \lambda_N/G$ and $\mathrm{LRM}_g(J_N, \widehat{w}) \ge J_N \widehat{w}_g - 1$, so, eventually,

$$n_{\min} \ge \frac{N\lambda_N}{2G}. \tag{54}$$

Step 2: conditional CLT for the leading influence term. Conditional on the pilot, the main samples are independent and their counts are fixed. Let

$$Z_N := \sum_g \frac{1}{n_g} \sum_{i=1}^{n_g} \phi_g(W_{g,i})$$

be the leading influence term. If $\sigma_g = 0$, then centering and zero variance imply $\phi_g = 0$ almost surely. Let $\mathcal{A} := \{g : \sigma_g > 0\}$. This set is nonempty because $S_\sigma > 0$, and $w_{\min}^* := \min_{g\in\mathcal{A}} w_g^* > 0$ because $\mathcal{A}$ is finite. By equation 53, the pilot event

$$B_N := \left\{ \min_{g\in\mathcal{A}} \frac{n_g}{N} \ge \frac{w_{\min}^*}{2} \right\}$$

satisfies $\mathbb{P}(B_N) \to 1$. Conditional on a pilot in B_N, the summands of $\sqrt{N}Z_N$ are independent and centered. Because the finite collection of influences is bounded, there is a deterministic $M_3 < \infty$ with $\mathbb{E}|\phi_g|^3 \le M_3$ for every g, and

$$\sum_{g\in\mathcal{A}}\sum_{i=1}^{n_g}\mathbb{E}\left[\left|\frac{\sqrt{N}}{n_g}\phi_g(W_{g,i})\right|^3\middle|\ \text{pilot}\right] \le M_3 N^{3/2}\sum_{g\in\mathcal{A}}\frac{1}{n_g^2} \le \frac{4M_3|\mathcal{A}|}{(w^*_{\min})^2\sqrt{N}} \longrightarrow 0.$$

The conditional variance is

$$N\sum_{g\in\mathcal{A}}\frac{\sigma_g^2}{n_g} \xrightarrow{p} \sum_{g\in\mathcal{A}}\frac{\sigma_g^2}{w_g^*} = S_\sigma^2 = V^* > 0.$$

Consequently, on an event whose pilot probability tends to one, the variance is bounded away from zero; dividing the preceding third-moment bound by its $3/2$ power verifies Lyapunov's condition, and hence Lindeberg's condition. The Lindeberg–Feller theorem applied conditionally on the pilot (van der Vaart, 1998, Proposition 2.27) therefore makes the conditional characteristic function of $\sqrt{N}Z_N$ converge in probability to that of $\mathcal{N}(0, V^*)$. Characteristic functions are bounded by one, so taking expectations over the pilot yields the unconditional convergence $\sqrt{N}Z_N \Rightarrow \mathcal{N}(0, V^*)$.

Step 3: nonlinear fixed-point remainder. The fixed-point remainder also remains negligible. Conditional on the pilot, Lemma 2 gives $\mathbb{E}[\|R_N\|^2 \mid \text{pilot}] \le Cn_{\min}^{-2}$ with deterministic C. By equation 54,

$$\sqrt{N}\|R_N\| = O_p\Big((\sqrt{N}\lambda_N)^{-1}\Big) = o_p(1), \qquad N\mathbb{E}\|R_N\|^2 = O\big((N\lambda_N^2)^{-1}\big) = o(1). \tag{55}$$

Together with the quantile-cell argument below, this proves the CLT.

Step 4: uniform integrability of the leading variance. For the normalized MSE, independence and centering give

$$\mathbb{E}[NZ_N^2 \mid \text{pilot}] = N\sum_{g:\sigma_g>0}\frac{\sigma_g^2}{n_g}. \tag{56}$$

For a group with $\sigma_g > 0$, let $A_{g,N} := \{\widehat{\sigma}_g \ge \sigma_g/2\}$. Since $\sum_j \widehat{\sigma}_j \le GB_\sigma$ and eventually $1 - \lambda_N \ge 1/2$, on $A_{g,N}$, $\widehat{w}_g \ge \sigma_g/(4GB_\sigma)$. On $A_{g,N}^c$, $\widehat{w}_g \ge \lambda_N/G$. Also $n_g \ge J_N\widehat{w}_g$ and eventually $N/J_N \le 2$, so $N/n_g \le 2/\widehat{w}_g$. Lemma 3 gives

$$\mathbb{E}\left[\frac{N}{n_g}\mathbf{1}_{A_{g,N}^c}\right] \le \frac{2GC_g}{\lambda_N}e^{-c_g m_N} \longrightarrow 0,$$

because $\log(1/\lambda_N) = o(m_N)$. On $A_{g,N}$, N/n_g is uniformly bounded and converges in probability to $1/w_g^*$. Hence equation 56 converges in expectation to V^*.

Step 5: wrong VaR cells and conclusion. Finally, conditional on the pilot, equation 34 and equation 54 bound the wrong-VaR-cell contribution to the normalized MSE by $CN\exp(-cN\lambda_N m_\alpha^2) = o(1)$. Cauchy–Schwarz removes the cross term with the L^2-negligible remainder. Thus

$$\sqrt{N}(\widehat{C}_N^{\text{TIS}} - C_{\alpha,K}) \Rightarrow \mathcal{N}(0, V^*), \qquad N\mathbb{E}[(\widehat{C}_N^{\text{TIS}} - C_{\alpha,K})^2] \to V^*,$$

which completes the proof. □

D.4 Anchored design: variance safeguard and asymptotic MSE

For the anchor in Equation 10, we define the occupancy scores, prove the component-relative variance safeguard, and establish Corollary 1.

Let $\widehat{\mu}_h(s)$ denote the state occupancy induced from the root by the fixed policy and the pilot transition estimate, with h steps remaining. In the stationary shared-kernel model, define

$$\widehat{o}_{s,a} := \sum_{h:(h,s)\in\mathcal{B}}\widehat{\mu}_h(s)\pi_h(a \mid s). \tag{57}$$

More generally, when a query group feeds several Bellman rows with known mixture coefficients, $\widehat{o}_g$ is the sum of the corresponding pilot-model row visitation probabilities times those coefficients. In the untied policy-mixture case $b = (h, s)$, this reduces to $\widehat{o}_b = \widehat{\mu}_h(s)$.

Let

$$p_g^{\mathrm{inf}} = \frac{\widehat{\sigma}_g}{\sum_j \widehat{\sigma}_j}, \qquad p_g^{\mathrm{occ}} = \frac{\widehat{o}_g}{\sum_j \widehat{o}_j}, \qquad p_g^{\mathrm{anc}} = \tfrac{1}{2}p_g^{\mathrm{inf}} + \tfrac{1}{2}p_g^{\mathrm{occ}},$$

using the uniform branch for p^{inf} if every estimated influence scale is zero. The occupancy denominator is positive because the retained rows include the root and $H \geq 1$. Apply the same exploration floor to each component, $w^x = (1-\lambda)p^x + \lambda \mathbf{1}/G$ for $x \in \{\mathrm{inf}, \mathrm{occ}, \mathrm{anc}\}$.

Proposition 4 (Component-relative safeguard)**.** *For the positive fractional designs before integer rounding,*

$$V(w^{\mathrm{anc}}) \leq 2\min\{V(w^{\mathrm{inf}}), V(w^{\mathrm{occ}})\}. \tag{58}$$

Proof. The common floor gives $w^{\mathrm{anc}} = (w^{\mathrm{inf}} + w^{\mathrm{occ}})/2$, so for every group

$$w_g^{\mathrm{anc}} \geq \tfrac{1}{2}w_g^{\mathrm{inf}}, \qquad w_g^{\mathrm{anc}} \geq \tfrac{1}{2}w_g^{\mathrm{occ}}.$$

Because $V(w) = \sum_g \sigma_g^2/w_g$ is decreasing in each coordinate separately,

$$V(w^{\mathrm{anc}}) \leq 2V(w^{\mathrm{inf}}), \qquad V(w^{\mathrm{anc}}) \leq 2V(w^{\mathrm{occ}}),$$

which proves the claim. □

Choice of mixture weight. For occupancy weight $\beta \in (0,1)$, the same argument gives

$$V\big((1-\beta)w^{\mathrm{inf}} + \beta w^{\mathrm{occ}}\big) \leq \min\left\{\frac{V(w^{\mathrm{inf}})}{1-\beta}, \frac{V(w^{\mathrm{occ}})}{\beta}\right\}.$$

The worst-case factor in this bound relative to the better component is $\max\{(1-\beta)^{-1}, \beta^{-1}\}$, minimized at $\beta = 1/2$; optimal finite-budget weights may differ.

This safeguard compares the two component designs before rounding; it does not bound finite-budget MSE. Corollary 1 identifies the limiting MSE constant of the full anchored estimator, including pilot cost and rounding. Finite-budget behavior is assessed in the workflow experiments (Sections E.2.6–E.10).

Corollary 1 (Asymptotic efficiency cost of anchoring)**.** *Under the assumptions and schedules of Theorem 3, let $v_g = o_g/\sum_j o_j$ be the population occupancy shares and set $a_g^* = (w_g^* + v_g)/2$. Using Equation 10 in Algorithm 1 gives*

$$\begin{aligned} \sqrt{N}\big(\widehat{C}_N^{\mathrm{anc}} - C_{\alpha,K}\big) &\Rightarrow \mathcal{N}(0, V_{\mathrm{anc}}), \qquad V^* \leq V_{\mathrm{anc}} \leq 2V^*. \\ N\mathbb{E}\big[(\widehat{C}_N^{\mathrm{anc}} - C_{\alpha,K})^2\big] &\to V_{\mathrm{anc}} = \sum_{g:\sigma_g>0} \frac{\sigma_g^2}{a_g^*}. \end{aligned} \tag{59}$$

Proof of Corollary 1. **Idea.** The occupancy shares are consistent, and the anchor retains at least half of every learned influence share. The proof of Theorem 3 therefore continues to control rare underallocation, with a changed limiting design.

Step 1: limiting shares and minimum counts. In fixed dimension, the pilot transition probabilities converge in probability to their population values. Finite-horizon occupancies are continuous functions of those probabilities, and their sum is positive. Thus the normalized pilot occupancy shares converge to v. The common floor vanishes, while Lemma 3 gives $w^{\mathrm{inf}} \to w^*$ in probability. Consequently,

$$w_g^{\mathrm{anc}} \xrightarrow{p} a_g^* = \tfrac{1}{2}(w_g^* + v_g).$$

For every active group $\sigma_g > 0$, $a_g^* \geq w_g^*/2 > 0$. Both floored components have weights at least λ_N/G, so the anchored counts obey the same lower bound $n_{\min} \geq N\lambda_N/(2G)$ eventually as in Equation 54. Rounding and the vanishing pilot fraction give $n_g^{\mathrm{anc}}/N \to a_g^*$ in probability.

Step 2: the influence term and its second moment. Conditional on the pilot, the main observations are independent. The bounded-influence conditional CLT used in Theorem 3 gives a Gaussian limit with variance V_{anc}; zero-scale groups contribute nothing. To justify convergence of second moments,

couple the hypothetical plain and anchored allocations to the same pilot. Write $J_N = N - Gm_N - 2G$, and let $n_g^{\inf}$ be the plain count. Largest-remainder rounding gives

$$n_g^{\mathrm{anc}} \geq 1 + J_N w_g^{\mathrm{anc}} \geq 1 + \tfrac{1}{2} J_N w_g^{\inf} \geq \tfrac{1}{2}(n_g^{\inf} - 1) \geq \tfrac{1}{4} n_g^{\inf},$$

where $n_g^{\inf} \leq 3 + J_N w_g^{\inf}$ and $n_g^{\inf} \geq 2$ were used. Hence

$$0 \leq N \sum_g \frac{\sigma_g^2}{n_g^{\mathrm{anc}}} \leq 4N \sum_g \frac{\sigma_g^2}{n_g^{\inf}}.$$

To make the moment transfer explicit, set $Y_N^{\inf} := N \sum_g \sigma_g^2 / n_g^{\inf}$ and $Y_N^{\mathrm{anc}} := N \sum_g \sigma_g^2 / n_g^{\mathrm{anc}}$. Step 4 of Theorem 3 gives $Y_N^{\inf} \to V^*$ in probability and $\mathbb{E} Y_N^{\inf} \to V^*$. Since these variables are nonnegative, this implies $Y_N^{\inf} \to V^*$ in L^1 and hence uniform integrability. The domination $0 \leq Y_N^{\mathrm{anc}} \leq 4 Y_N^{\inf}$ therefore makes $\{Y_N^{\mathrm{anc}}\}$ uniformly integrable as well. Because $Y_N^{\mathrm{anc}} \to V_{\mathrm{anc}}$ in probability by Step 1, uniform integrability yields $\mathbb{E} Y_N^{\mathrm{anc}} \to V_{\mathrm{anc}}$. This proves the normalized-MSE limit for the leading influence term.

Step 3: remainder, quantile, and efficiency cost. The common minimum-count bound gives the same vanishing normalized second moment of the Bellman remainder and the same exponentially small wrong-quantile contribution as in Theorem 3. The cross term vanishes by Cauchy–Schwarz. The CLT and MSE limit thus hold for the full CVaR estimator. Finally, $a_g^* \geq w_g^*/2$ on active groups implies

$$V_{\mathrm{anc}} \leq 2 \sum_{g:\sigma_g > 0} \frac{\sigma_g^2}{w_g^*} = 2V^*.$$

Cauchy–Schwarz gives $V_{\mathrm{anc}} \geq V^*$ for any probability allocation, including allocations with zero weights only on zero-influence groups. □

D.5 When learning an allocation repays its pilot

The variance identity in Equation 51 separates allocation opportunity from learning error. It remains valid when some influences vanish, provided $S_\sigma = \sum_g \sigma_g > 0$ and the evaluated design is positive. Let $n = N - Gm$, $\rho = Gm/N$, retain the oracle shares $w_g^* = \sigma_g / S_\sigma$ from Equation 6, and let $\widetilde{w}_g = n_g / n$ be the actual main-sample fractions after rounding. Define

$$D(w^* \| w) := \sum_g \frac{(w_g^* - w_g)^2}{w_g}.$$

Because both w^* and w sum to one,

$$D(w^* \| w) = \sum_g \frac{(w_g^*)^2}{w_g} - 1.$$

Conditional on the pilot, the centered leading influence term $L_N = \sum_g n_g^{-1} \sum_i \phi_g(W_{g,i})$ therefore satisfies

$$N \mathbb{E}[L_N^2 \mid \text{pilot}] = N \sum_g \frac{\sigma_g^2}{n_g} = \frac{N}{n} V^* \sum_g \frac{(w_g^*)^2}{\widetilde{w}_g} = \frac{V^*}{1-\rho} \{1 + D(w^* \| \widetilde{w})\}.$$

Averaging over the pilot gives the exact identity

$$N \mathbb{E}[L_N^2] = \frac{V^*}{1-\rho} \left\{1 + \mathbb{E} D(w^* \| \widetilde{w})\right\}. \tag{60}$$

Here the expectation on the right is over pilots. For a deterministic baseline using all N queries without a pilot and positive actual query fractions v, write $A_v = V(v)/V^*$. Its leading variance is $V(v)/N$. Learning improves on this benchmark at the level of the influence term exactly when

$$1 + \mathbb{E} D(w^* \| \widetilde{w}) < (1-\rho) A_v. \tag{61}$$

The three quantities have distinct roles: A_v is the available allocation advantage, D penalizes inaccurate shares, especially underallocation, and ρ charges the discarded pilot. For two learned methods with the same pilot size, the common factor $1/(1-\rho)$ cancels, so their comparison depends on their expected allocation penalties. These are identities for the linearized error, not finite-budget guarantees for the nonlinear CVaR estimator. Population influences are needed to evaluate them, so their use in the experiments is diagnostic rather than an operational rule for choosing a method.

E Approximation and Experimental Protocols

This section closes two gaps left by the asymptotic allocation theory. Appendix E.1 controls error from the categorical grid; the remaining subsections give the protocols and evidence supporting **Q1–Q5**, including the regimes where tail targeting helps, where it does not, and why small pilots can fail. Each primary protocol defines its query and charged budget.

E.1 Representation error

A fixed grid introduces error even with exact conditional laws. The following bound justifies the separation of approximation and sampling error at the end of Section 2. Let $Z_h^*(s) \sim \eta_h^*(s)$ denote the projected h-step return and $G_h(s)$ its true-return counterpart. For every state s,

$$|\operatorname{CVaR}_\alpha(\eta_H^*(s)) - \operatorname{CVaR}_\alpha(G_H(s))| \le H\Delta. \tag{62}$$

Proof. Realize each categorical projection as randomized rounding to adjacent grid atoms (Rowland et al., 2018): for $y \in [z_j, z_{j+1}]$, set $\widetilde{y} = z_j$ with probability $(z_{j+1} - y)/(z_{j+1} - z_j)$ and $\widetilde{y} = z_{j+1}$ otherwise; for $y \ge H$, set $\widetilde{y} = H$. Then $\mathcal{L}(\widetilde{y}) = \Pi_C \delta_y$ and $|\widetilde{y} - y| \le \Delta$ whenever $y \le H$ (inputs below zero do not arise here). Couple the projected and true return recursions using the same actions, rewards, next states, and these rounding variables. Conditional on a coupled next state, use the inductive coupling for the two continuation returns. If their difference is at most $(h-1)\Delta$, adding the same reward preserves that difference and adjacent-grid rounding adds at most Δ. If the projected pre-rounding value exceeds H, clipping it to H cannot increase its distance from the true h-step return, which lies in $[0, h] \subseteq [0, H]$. Starting from equal zero-step returns, induction yields

$$|Z_h^*(s) - G_h(s)| \le h\Delta \qquad \text{almost surely.} \tag{63}$$

If $|X - Y| \le c$ almost surely, then $F_X(x - c) \le F_Y(x) \le F_X(x + c)$ for every x, which implies $|F_X^{-1}(u) - F_Y^{-1}(u)| \le c$ for $u \in (0, 1)$. Integrating over $u \in (0, \alpha)$ and dividing by α gives Equation 62: the integration interval's length cancels the factor $1/\alpha$. □

E.2 Experimental Evidence and Protocols

The protocols below support **Q1**–**Q3** in Section 5; Section E.3 gives the held-out FinQA protocol for **Q4**. MSE is the average squared error over independent replications, with Monte Carlo SE equal to the sample standard deviation of squared errors divided by the square root of the replication count; bars show 1.96 SEs. Rollouts use $\lfloor N/H \rfloor$ full trajectories (leaving fewer than H transition slots unused), whereas conditional allocations exhaust N. Bold marks sample-only point minima, including displayed ties, without a superiority claim; population references are shaded. Uncertainty is conditional on the fixed tasks or panels.

Method key and common estimator. All conditional-query methods use the categorical CVaR plug-in. Fixed-score designs normalize scores, add the floor in Equation 8, and round to exhaust the budget; learned scores pay for and discard a pilot, whereas population references use exact laws without a pilot. **Uniform** assigns equal shares, and **learned occupancy** (**occup./Occ.**) uses pilot-model visit counts $\widehat{o}_g$ from Equation 57. **Learned mean** uses the pilot-estimated standard deviation of the mean-return influence; for a shared kernel,

$$\psi_{s,a}(W) = \sum_{h=1}^{H} \mu_h(s)\pi_h(a \mid s)\{R + v_{h-1}(S') - \mathbb{E}_{P_{s,a}}[R + v_{h-1}(S')]\},$$

where $\mu_h(s)$ is visitation with h steps left and $v_h(s) = \mathbb{E}[G_h(s)]$; an untied block uses $\mu_h(s)\,\mathrm{sd}_{P_b}(R + v_{h-1}(S'))$. **Occupancy (population)** and **mean influence (population)** use the corresponding exact scores, and **complete rollout** (**rollout/Roll.**) takes empirical CVaR of full fixed-policy returns (Thomas & Learned-Miller, 2019).

Plain/shared TIS uses $\widehat{\sigma}_g$; **anchored** TIS (**anchored/Anch.**) averages the tail and occupancy designs. **Oracle+floor** uses exact σ_g and is a population allocation reference, not a finite-budget MSE lower bound. **Reachability only** and **local scale only** replace $\sigma_b = d_b\tau_b$ by exact d_b or τ_b (Section C.3), separating threshold-dependent adjoint mass from local shortfall variability; d_b need not

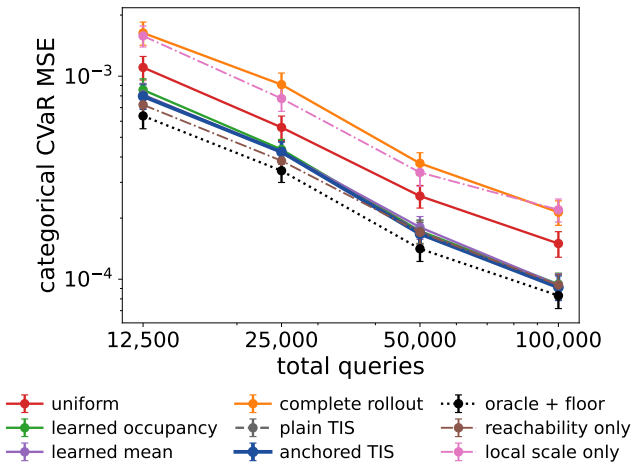


Figure 3: Controlled $H = 5$ MRP (21 untied blocks): CVaR MSE versus total queries for learned methods and population references; bars are 1.96 Monte Carlo SEs.

equal occupancy. TIS **no-cov** removes cross-layer covariance from the allocation score while retaining pooled estimation, with **oracle no-cov** its population analogue; **untied** TIS fits separate layer kernels under the same total budget. **MC-UCB (frozen)** sequentially allocates using uncertainty in pilot-frozen influence scores (Section E.2.4), and **pilot answer entropy** uses Shannon entropy of pilot answer frequencies, summed over confidence labels.

E.2.1 Controlled stochastic Markov reward process

This auxiliary multi-step check supports the mechanism behind **Q1–Q2**: in the untied setting, the oracle score combines propagation to the root and local shortfall variability. The $H = 5$ MRP has five states per layer, rewards in $\{0, 1/2, 1\}$, grid spacing $\Delta = 1/2$, and stochastic rewards and transitions in every block, giving $G = 21$ independently sampled layer–state blocks. At $\alpha = .1$, exact enumeration gives CVaR .9086716, margin .0106854, and $V_{\text{unif}}/V^* = 12.98416/7.53437 = 1.72332$.

We use total budgets 12,500–100,000, the common $m_N = \lceil 4N^{2/3}/G \rceil$ discarded pilot and $\lambda_N = N^{-1/4}$ floor, and 400 replications. At $N = 100{,}000$, TIS spends 8,631 pilot queries and reaches .615 of uniform MSE; anchored TIS, learned mean, and learned occupancy are close, while rollouts are worse than uniform (Figure 3).

Across 50 perturbed stochastic instances (Dirichlet seeds 100–149), the median V_{unif}/V^* at $\alpha = .05, .1, .2$ is $1.95344, 1.92311, 1.82727$; ratios span 1.34558–3.56479 over all instance–risk pairs, with positive margins throughout. This is a population-level breadth check.

E.2.2 Learned-budget runs for the controlled separation

This experiment directly supports **Q1–Q2**: it asks whether a charged pilot learns the population separation of Equation 44. We use $G = 10$, $\alpha = .1$, and $t \in \{0, \frac{1}{2}, 1\}$ under a protocol frozen before simulation. Analytically, CVaR is .3, the margin is .09, occupancy-to-oracle variance ratios are 1, 1.29971, and 10, and conditional moments and the root law agree across t to 7×10^{-17}.

Each cell uses 1,000 replications at 25, 50, 100, 200, and 400 queries per kernel; pilot size, floor, rounding, and fallback follow the common protocol, and rollouts use the same transition budget on an independent stream. Because the one-step policy fixes occupancy at $1/G$, uniform is also the no-pilot occupancy design; **occupancy + pilot** discards the common pilot before using that same allocation, isolating pilot cost. At 100/400 queries per kernel the pilot consumes 40/101 draws per group ($40\%/25.25\%$); learned mean, TIS, and the anchor pay the same cost, whereas uniform, rollouts, and oracle+floor do not. Rollouts randomize actions while uniform fixes equal counts. At $t = 0$, each group's zero-reward probability is .01, so the pilots miss it with probability .669/.362

Table 3: Absolute MSE $\times 10^5$ (Monte Carlo SE) for the six main controlled-separation settings; 1,000 replications per entry.

	$t=0$		$t=\frac{1}{2}$		$t=1$	
Queries/kernel	100	400	100	400	100	400
Uniform	**11.07** (0.48)	2.98 (0.13)	11.44 (0.50)	**2.69** (0.12)	9.79 (0.44)	2.46 (0.12)
Occupancy + pilot	17.68 (0.79)	3.89 (0.17)	19.12 (0.86)	3.73 (0.17)	16.51 (0.76)	3.22 (0.16)
Learned mean	18.05 (0.79)	3.82 (0.16)	19.53 (0.92)	3.74 (0.16)	18.32 (1.07)	3.30 (0.17)
Complete rollout	11.16 (0.56)	**2.84** (0.12)	**11.27** (0.50)	2.87 (0.12)	10.84 (0.46)	2.90 (0.13)
TIS	59.96 (3.19)	12.82 (0.66)	41.45 (2.68)	8.47 (0.44)	**2.17** (0.11)	**0.40** (0.02)
anchored TIS	22.55 (0.98)	4.52 (0.20)	19.02 (0.97)	3.14 (0.14)	3.65 (0.17)	0.66 (0.03)
Oracle + floor	11.07 (0.48)	2.98 (0.13)	8.62 (0.42)	2.08 (0.09)	1.26 (0.06)	0.30 (0.01)

Table 4: Inventory MSE $\times 10^3$ over 300 replications versus charged queries per retained block.

Method	150	300	600	1,200
Uniform	1.232	0.583	0.303	0.176
Reachability only	1.091	0.659	0.297	0.177
Local scale only	1.908	1.049	0.531	0.285
Oracle + floor	0.731	0.329	0.150	0.100
Learned occupancy	1.517	0.634	0.311	0.161
Learned mean	1.361	0.571	0.302	0.152
Complete rollout	2.794	1.774	0.836	0.391
anchored TIS	1.012	0.456	0.217	0.137
TIS	**1.007**	**0.414**	**0.174**	**0.116**

at 100/400 queries, explaining why learning can hurt when uniform is optimal. At $t = 1$, the median TIS share of the informative kernel rises from .62 to .88 across budgets (oracle share 1); its 10th percentile is .10–.75 at 25–50 queries, exposing the underallocation that anchoring mitigates. Table 3 reports absolute MSEs and Monte Carlo SEs for the main-table cells.

E.2.3 Seasonal base-stock inventory evaluation

This stage-dependent benchmark supports **Q2**; each (h, s) is a separate query group (the untied case of Section A). Inventory is $\{0, \dots, 6\}$, a fixed seasonal policy orders toward four or five units over $H = 8$, demand is truncated Poisson with seasonally varying mean, and a disruption with probability .06 (.08 at capacity) removes one extra unit and lowers the reward category. Sales, ordering, holding, and lost-demand terms determine profit, quantized to $\{0, 1/2, 1\}$ on grid $\Delta = 1/2$.

From zero initial inventory at $\alpha = .1$, backward structural reachability leaves $B = 41$ blocks. Exact enumeration gives categorical CVaR 1.1512371, margin 0.0436202, and

$$V_{\text{unif}} = 7.54133, \qquad V^* = 4.13541, \qquad V_{\text{unif}}/V^* = 1.82360.$$

Budgets are 150, 300, 600, and 1,200 queries per retained block (6,150–49,200 total calls); pilot, floor, rounding, and sample splitting follow the controlled-MRP protocol. We use 300 replications.

Pilot fractions are $22.0\%, 17.3\%, 13.8\%, 10.9\%$. TIS has the lowest observed learned-method MSE at every budget (Table 4); at 1,200 queries per block it is 15.4% above oracle+floor. Rollouts exceed uniform throughout: with $H = 8$, 49,200 transitions yield only 6,150 returns, or 615 returns' worth of mass in the worst decile.

At 600 queries per block, 300 independent pilots per setting isolate allocation learning through $V(\widehat{w})/V^*$; this excludes the main-sample cost of larger pilots, for which the leading cost-inclusive MSE is $V(\widehat{w})/(N - Gm_N)$. Table 5 gives medians and 90th percentiles; smaller exploration exponents correspond to larger floors.

E.2.4 Slippery CliffWalking with stationary state–action kernels

Figure 2(a) provides the shared-kernel **Q2** benchmark: stationary state–action laws are reused across Bellman stages. We use Gymnasium slippery CliffWalking-v1 (Towers et al., 2025), a 4×12 grid

Table 5: Inventory pilot sensitivity: median (90th percentile) $V(\widehat{w})/V^*$ over 300 pilots.

Pilot fraction	$\lambda = N^{-1/6}$	$N^{-1/4}$	$N^{-1/3}$
3.7%	1.124 (1.184)	**1.111 (1.184)**	1.123 (1.212)
7.2%	1.082 (1.108)	1.057 (1.087)	**1.054 (1.088)**
14.2%	1.064 (1.073)	1.034 (1.043)	**1.028 (1.038)**
28.3%	1.055 (1.060)	1.023 (1.028)	**1.015 (1.020)**

Table 6: CliffWalking MSE $\times 10^3$ over 500 replications versus charged queries per reachable kernel.

Method	50	100	200	400
Uniform	1683.126	818.420	363.617	216.877
Occupancy (population)	26.269	12.474	5.727	2.741
Mean influence (population)	19.182	9.731	4.371	2.164
Learned mean	**27.413**	11.367	5.349	2.307
Learned occupancy	35.792	15.924	6.450	3.189
Complete rollout	65.883	32.221	16.257	7.954
MC-UCB (frozen)	1863.426	840.730	351.955	171.316
Oracle no-cov	16.203	7.462	3.416	1.655
Oracle + floor	16.288	7.631	3.434	1.674
TIS no-cov	29.979	10.821	**4.242**	**1.875**
anchored TIS	28.427	11.551	5.362	2.182
TIS	29.560	**10.755**	4.305	1.884

with start 36, goal 47, cliff cells 37–46, and four actions; each action realizes its intended or either perpendicular direction with probability $1/3$, boundary moves stay put, cliffs reset to start, and the fixed-horizon wrapper makes the goal absorbing.

We set $H = 20$ and map raw rewards $-100, -1, 0$ to $0, .99, 1$. With zero reward after termination, the normalized return is exactly $20 + G_{\text{episodic}}/100$, so the transform preserves lower-tail ordering and CVaR. The grid contains every sum of 20 elements of $\{0, .99, 1\}$ (231 atoms), hence the Bellman recursion is exact. The primary policy is $\epsilon = .05$-soft around a route crossing row 2 immediately above the cliff; a prespecified safe diagnostic policy crosses row 1, and off-route states first steer toward the chosen corridor. Structural reachability leaves 149 state–action groups, including the absorbing goal. For the primary policy at $\alpha = .1$, CVaR is 11.8709354, the margin is 0.0139085, and $V_{\text{unif}}/V^* = 84.6973$; across both policies and $\alpha \in \{.05, .1, .2\}$, exact ratios span 77.5748–116.3908. Their minimum route lengths are 13 and 15, so $H = 20$ allows completion and slippery deviations with a tractable exact grid. All reported methods evaluate the primary policy; the safe policy is a separate policy–risk check.

Population occupancy and mean-influence references use exact scores with the same $N^{-1/4}$ floor; learned counterparts use the pilot fit. The discarded pilot is $m_N = \max\{8, \lceil 4N^{2/3}/149 \rceil\}$ per group. For MC-UCB (Carpentier et al., 2015), whose original guarantee concerns fixed-stratum weighted means rather than our Bellman estimator, we freeze the pilot influence scores and treat fresh main draws as bounded score arms. After two draws per group it selects the largest

$$B_{g,t} = \frac{1/G}{T_{g,t-1}} \left(\widehat{s}_{g,t-1} + \frac{2\beta}{\sqrt{T_{g,t-1}}} \right),$$

where $T_{g,t-1}$ is the main-sample count and $\widehat{s}_{g,t-1}$ the score standard deviation. We use the published bounded-arm choices $\delta = n^{-9/2}$ and $\beta = c\sqrt{\log(2/\delta)}$, with n the main budget and c the largest frozen-arm range. Positive affine rescaling leaves the rule unchanged; computing c uses the public three-outcome support but not its probabilities, which is extra information relative to unknown-support simulators. MC-UCB and TIS share the charged pilot, exhaust the same main budget, use the same categorical plug-in, and select no hyperparameter from outcomes. We use 500 independent replications. The no-covariance ablations in Table 6 show that most gains here come from pooling reused kernels; the theoretically required cross-layer covariance has only a small numerical effect.

Pilot fractions are $22.0\%, 17.0\%, 13.0\%, 10.25\%$. At 50 queries per kernel, TIS is resolved worse than population mean influence and oracle+floor, while its differences from population occupancy and learned mean are unresolved; it beats population occupancy from 100 queries and population mean at 400. At 400, MSE reductions are 31.3%, 12.9%, and 18.3% versus population occupancy, population mean, and learned mean. Deleting covariance from the learned score yields no resolved difference at any budget. Rollouts beat uniform but trail the occupancy and influence designs, except MC-UCB. MC-UCB's mean realized first-order variance ratios are $96.69, 85.80, 76.18, 67.77$ versus uniform's 84.70; under its published significance schedule, the confidence bonus dominates at these pull counts, keeping allocation near uniform and yielding only modest improvement at larger budgets.

E.2.5 Additional public stationary Gymnasium environments

These breadth checks test whether the shared-kernel conclusions extend beyond CliffWalking and expose a regime in which a smoother mean score can be preferable. We use stationary `FrozenLake-v1` and rainy `Taxi-v4` (Towers et al., 2025), with fixed policies, reused state kernels, the CliffWalking budgets and estimator, and prespecified screening for at least two stochastic reachable kernels, nonzero oracle influence variance, a positive categorical margin, and an exact finite grid.

FrozenLake. The slippery 8×8 task leaves 50 reachable nonterminal kernels under a fixed success-maximizing policy. Its failure probability $.13704$ makes the first-order tail signal zero at $\alpha = .05, .1$, so $\alpha = .2$ is the smallest prespecified passing level; there CVaR is $.3147769$, the margin is $.0629554$, and $V_{\text{unif}}/V^* = 3.92589$. **Rainy Taxi.** The fixed shortest-route policy leaves 37 reachable nonterminal kernels. The positive affine reward transform $(r+1)/21$ gives the exact grid $\{0, \ldots, 220\}/21$ and preserves CVaR; at $\alpha = .1$, transformed CVaR is 9.0541409, the margin is $.0121392$, and $V_{\text{unif}}/V^* = 2.33938$.

Both tasks use 300 replications at 50–400 queries per reachable kernel. At 400 queries, TIS MSE is $.00791\,(.00079)$ on FrozenLake and $.00020\,(.00002)$ on Taxi, versus uniform $.01464\,(.00119)$ and $.00033\,(.00003)$. Learned mean is better on FrozenLake ($.00577\,(.00051)$) and tied at displayed precision on Taxi ($.00020\,(.00002)$); differences from population occupancy and covariance-deleted TIS are unresolved, while oracle+floor is best on both. At smaller budgets, pilot cost can make TIS worse than uniform or population references. These checks reinforce the structural limit rather than a universal win: when tail and mean signals align, the smoother mean score can be as good as or better than tail targeting.

E.2.6 Structured language-model workflow evaluation

Table 7: **MMLU-Pro: anchoring mitigates plain TIS's observed failures.** Panel MSE/uniform MSE at $H = 6$, $\alpha = .1$, 400 queries/kernel.

Generator	TIS	occup.	rollout	**anchored**	oracle +floor
Qwen3-4B	1.38	.090	.074	**.057**	.058
Phi-4-mini	.49	.113	.072	**.066**	.025
Granite-4.2-8B	.15	.126	.107	**.070**	.034
Mistral-24B	.21	.138	.127	**.093**	.055
Qwen3-32B	.41	.079	**.055**	.056	.028
GLM-4-32B	1.77	.087	.069	**.065**	.034

These frozen-law experiments support **Q3**: exact targets diagnose plain-TIS pilot failures and the effect of anchoring. We use 50 stratified ten-option MMLU-Pro questions (Wang et al., 2024). Qwen3-4B-Instruct-2507 is primary (Qwen Team, 2025b); Phi-4-mini-instruct and Granite-4.2-8B are cross-family checks (Microsoft et al., 2025; Granite Team, IBM, 2026); follow-up generators (†) are Mistral-Small-24B-Instruct-2501, Qwen3-32B (thinking disabled), and GLM-4-32B-0414 (Mistral AI, 2025; Qwen Team, 2025a; Zhipu AI, 2025). All use the same panel, prompts, decoder, policy, budgets, and methods; floor, utility, panel, and policy sensitivities are post hoc.

Each question is a separate finite-horizon Markov reward process. From $S_0 = \varnothing$, state $S_t = (j_t, c_t)$ records the latest answer $j \in \{1, \ldots, 10\}$ and confidence $c \in \{.1, \ldots, .9\}$. The root action is `solve`; reviews use `reconsider` for $c \leq .3$, `challenge` for $.4 \leq c \leq .6$, and `verify` for $c \geq .7$. The controller tracks $h = H - t$, but prompts omit history and stage index, so a fixed state and action have the same next-response law at every stage. Thus the pair is Markov, the dynamics are stationary, and any retained prompt can be queried directly rather than reached by rollout.

Table 8: MMLU-Pro and FinQA workflow models; both use the confidence-band review policy.

Component	MMLU-Pro	FinQA
State after a call	Answer j and confidence c (10×9 possibilities)	Candidate i and confidence c (8×9 possibilities)
Initial action	`solve`	`select`
Number of calls	$H = 2, 4, 6$, including the initial solve	$H = 3$: selection and two reviews
Stage index	$h = H - t$ calls remaining; stop at $h = 0$	$h = 3, 2, 1, 0$; stop at $h = 0$
Query groups	$1 + 90 = 91$ per question	$1 + 72 = 73$ per question and workflow
Reward on $s \to s'$	$r_q(s')$: normalized Brier utility	$[u(s') - u(s) + 1]/2$: shifted utility change
Total return	$\sum_{t=1}^{H} r_q(S_t)$	$[H + u(S_H)]/2$, with $u(S_0) = 0$

Response probabilities factor as restricted answer softmax times conditional confidence softmax, with each label one token. A query to $g = (q, s, a)$ draws S' and reward $r_q(S')$:

$$P_{q,s,a}(r, s') = P_{\text{LLM}}(s' \mid \text{prompt}(q, s, a))\, \mathbf{1}\{r = r_q(s')\}. \tag{64}$$

For example, $(1, .2)$ selects `reconsider`; response $(3, .8)$ earns $r_q(3, .8)$ and, if a call remains, selects `verify`. One fitted law and sample count are shared across all uses of g, but its return effect depends on calls remaining; TIS therefore sums these stage effects before computing the group scale. Allocation is learned separately for every question, generator, and H. With one action per state there are 91 groups per question (4,550 per generator).

Reward and exact target. For response $y = (j, c)$, define the ten-class forecast and normalized Brier utility (Brier, 1950)

$$p_k(y) = \begin{cases} c, & k = j, \\ (1 - c)/9, & k \neq j, \end{cases} \qquad r_q(y) = 1 - \frac{1}{2} \sum_{k=1}^{10} \left(p_k(y) - \mathbf{1}\{k = j_q^*\} \right)^2, \tag{65}$$

where j_q^* is the published correct option; no learned judge is used. The return $G_{H,q} = \sum_{t=1}^{H} r_q(S_t)$ includes the initial answer and every review, measuring cumulative response utility (FinQA instead uses terminal severity). Dividing by H rescales CVaR and MSE but leaves within-horizon ratios and allocations unchanged.

The grid is the union of all attainable partial-return supports for $h = 0, \ldots, H$ plus endpoints $0, H$: 121, 617, and 983 atoms at $H = 2, 4, 6$. Closure makes both the categorical recursion and stop-loss interpolation exact; each question–horizon pair has its own target, margin, and scales. We use $\alpha = .1$ primarily and .2 as a sensitivity check. Partial returns are essential: two deterministic .5 rewards have true sum 1, but backing them up on $\{0, 1, 2\}$ yields $\frac{1}{4}\delta_0 + \frac{1}{2}\delta_1 + \frac{1}{4}\delta_2$, preserving the mean while driving $\text{CVaR}_{.1}$ to 0; including .5 restores exactness. Across all 1,800 question–horizon–risk cells, full-horizon-only grids have median/90th-percentile absolute CVaR gaps $.02/.35$, whereas closed-grid gaps are below 10^{-14}. All reported runs use the closed grids; grid smoothing partly masks the primary generator's depth failure.

Ground truth and validation. Enumerating each prompt's 90 probabilities and applying dynamic programming gives exact targets, influences, and population references; sample-only methods receive fresh (R, S') draws. Enumeration is practical here only because the output alphabet is restricted, so the study measures logical-query efficiency under frozen laws while remaining relevant to longer outputs, restricted APIs, and stochastic tools. All kernels pass the designated direct-generation audit. Methods otherwise follow Section E.2; MC-UCB uses its sequential index.

Budgets and uncertainty. Each generator–horizon cell has $J = 300$ replications at $b = 50, 100, 200, 400$ queries per group, so $N = 91b = 4{,}550, 9{,}100, 18{,}200, 36{,}400$ draws per question. The charged pilot $m_N = \max\{8, \lceil 4N^{2/3}/91 \rceil\}$ uses $13, 20, 31, 49$ draws per group and

Table 9: MMLU-Pro, $\alpha = .1$, 400 queries/shared kernel: MSE/uniform MSE. † marks follow-up generators; Table 10 gives absolute TIS MSE.

Generator	H	plain TIS	anchored TIS	learned mean	learned occup.	TIS rollout	TIS no cov.	untied	MC-UCB	oracle +floor
Qwen3-4B	2	.719	.030	.307	.036	**.028**	.719	.724	1.04	.192
	4	1.02	.059	.707	.069	**.057**	1.02	4.28	1.08	.058
	6	1.38	**.057**	1.00	.090	.074	1.35	7.67	1.06	.058
Phi-4-mini	2	.181	.027	.106	.032	**.023**	.181	.183	1.03	.015
	4	.326	**.047**	.264	.074	.048	.327	1.33	1.11	.020
	6	.487	**.066**	.425	.113	.072	.493	3.00	1.48	.025
Granite-4.2-8B	2	.172	.033	.038	.042	**.029**	.172	.160	1.06	.021
	4	.127	**.051**	.065	.079	.062	.127	.834	.981	.027
	6	.154	**.070**	.089	.126	.107	.164	1.56	1.09	.034
Mistral-24B†	2	.140	.038	.041	.043	**.026**	.140	.123	.967	.020
	4	.148	**.067**	.100	.089	.075	.144	.763	1.07	.038
	6	.214	**.093**	.159	.138	.127	.221	1.90	1.13	.055
Qwen3-32B†	2	.321	.028	.199	.030	**.023**	.321	.308	.956	.108
	4	.410	**.044**	.373	.058	**.044**	.409	2.43	1.03	.036
	6	.411	.056	.440	.079	**.055**	.412	3.71	1.08	.028
GLM-4-32B†	2	.438	.035	.426	.039	**.025**	.438	.425	1.06	.027
	4	1.08	.062	.790	.079	**.055**	1.08	4.24	1.28	.042
	6	1.77	**.065**	1.42	.087	.069	1.80	6.95	1.29	.034

$\lambda_N = N^{-1/4}$; the remaining fresh draws follow the learned shares and rounding rule, and the pilot is excluded from the final estimate. One logical query is a constrained two-token draw; ground-truth enumeration is outside this budget. Fixed-allocation methods share outcome streams, sequential methods use separate streams, and replications are independent. For squared-error difference $D_{q,\ell}$ between methods A and B on question q and replication ℓ,

$$D_\ell = \frac{1}{Q}\sum_{q=1}^{Q} D_{q,\ell}, \qquad \widehat{\Delta} = \frac{1}{J}\sum_{\ell=1}^{J} D_\ell, \qquad \widehat{\mathrm{se}}(\widehat{\Delta}) = \frac{s_D}{\sqrt{J}}, \tag{66}$$

where $s_D^2 = (J-1)^{-1}\sum_\ell (D_\ell - \widehat{\Delta})^2$ and $z = \widehat{\Delta}/\widehat{\mathrm{se}}(\widehat{\Delta})$. This fixed-panel calculation allows arbitrary within-replication dependence among questions; uncertainty is simulation error conditional on the panel, not population generalization.

Checks and main pattern. All 150 question–horizon margins are positive at each risk level, and shared and population-untied targets agree within 10^{-10}. Median population V_{unif}/V^* at $\alpha = .1$ is $46, 52, 37, 27, 54, 42$ in table order. Table 9 shows occupancy below plain TIS in all 18 settings and the anchor below both; rollouts have the lowest sample-only point MSE for all generators at $H = 2$, while the anchor does so for five at $H = 6$. MC-UCB gives no consistent gain over uniform (.956–1.48), population occupancy/mean are strong (.024–.103/.016–.074), and answer entropy is poor (2.09–7.74). Pooling matters more than covariance correction: deeper untied evaluation splits data across $1 + 90(H-1)$ pilot groups, whereas deleting covariance changes little (population no-cov is within .003 of oracle+floor in uniform-MSE units). Tables 11–12 give budget and risk sensitivity.

Depth failures and anchoring. At $H = 6$, plain TIS exceeds uniform MSE for Qwen3-4B (1.38) and GLM-4-32B (1.77), and more budget does not reliably remove the failures (Table 11); at $H = 4$ the full sweep reaches 1.10 and 1.21. At $\alpha = .2$, only GLM $H = 6$ remains above uniform (1.52), while its $H = 4$ difference is unresolved. For GLM's primary $H = 4, 6$ cells, bias is at most 3% of MSE but realized population-scale variance is 1.36 and 2.66 times uniform: rare pilots starve influential groups, making σ_g^2/w_g large (Section E.10), outside Proposition 3's local regime.

A retrospective floor sweep supports the underallocation diagnosis: at $H = 6$, $\alpha = .1$, 400 queries, increasing the floor from .072 to .40 lowers Qwen and GLM TIS/uniform MSE from $1.38/1.77$ to $.47/.55$, with still larger floors reducing them further. Because the sweep followed the failures, it is diagnostic rather than a tuning result; held-out FinQA below tests the subsequently frozen

Table 10: Absolute TIS panel MSE (Monte Carlo SE), $\alpha = .1$, 400 queries/kernel; 300 replications.

Generator	$H = 2$	$H = 4$	$H = 6$
Qwen3-4B	2.2e-04 (1e-05)	9.7e-04 (6e-05)	2.7e-03 (2e-04)
Phi-4-mini	2.8e-04 (1e-05)	1.7e-03 (1e-04)	5.1e-03 (5e-04)
Granite-4.2-8B	9.8e-05 (5e-06)	1.3e-04 (1e-05)	2.4e-04 (1e-05)
Mistral-24B†	1.2e-04 (2e-05)	3.5e-04 (3e-05)	9.6e-04 (1e-04)
Qwen3-32B†	5.7e-04 (4e-05)	2.9e-03 (3e-04)	7.5e-03 (8e-04)
GLM-4-32B†	4.1e-04 (3e-05)	4.2e-03 (2e-04)	1.9e-02 (9e-04)

Table 11: Budget sensitivity at $H = 6$, $\alpha = .1$: shared/untied TIS MSE relative to uniform at equal total budget.

	shared TIS				untied TIS			
Generator	50	100	200	400	50	100	200	400
Qwen3-4B	.819	.851	**1.22**	**1.38**	8.59	4.53	5.44	7.67
Phi-4-mini	.448	.499	.469	.487	4.29	2.02	2.55	3.00
Granite-4.2-8B	.469	.386	.248	.154	5.49	1.64	2.06	1.56
Mistral-24B†	.497	.462	.391	.214	5.81	1.58	2.08	1.90
Qwen3-32B†	.681	.606	.497	.411	5.45	2.71	2.91	3.71
GLM-4-32B†	**1.02**	**1.44**	**1.76**	**1.77**	6.39	2.65	4.81	6.95

anchor. Workflow selection on this MMLU panel is nearly saturated, so the informative outcome is estimation error rather than final pairwise choice.

Three post-hoc variations point to the same pilot-reliability mechanism. Under an asymmetric confidence-weighted utility, Qwen plain-TIS/uniform ratios at $H = 2, 4, 6$ are $.39, 1.02, 1.06$, versus anchor $.031, .053, .056$; on a disjoint high-stakes panel they are $.96, .42, .72$ versus anchor $.039, .052, .052$. A cautious policy again gives plain TIS 1.04 of uniform at $H = 6$ while occupancy is $.13$. These checks support the diagnosis only; none was used to choose the frozen anchor.

E.2.7 Cost-to-accuracy analysis and computation accounting

A lower MSE at one budget need not imply fewer queries at every target accuracy, so we convert error curves to query cost and report computation separately. These comparisons are retrospective (no accuracy tolerance was declared before the runs), and charged query counts include pilots and rollout transitions.

Multi-target relative cost. At 21 RMSE targets in the common attained range, we use log–log first-crossing interpolation without extrapolation; out-of-grid crossings retain budget bounds. Pointwise 5th/95th-percentile envelopes come from empirical bootstrap when replication rows are available (CliffWalking TIS/learned mean, inventory TIS/uniform, all FinQA pairs) and Gaussian sensitivity otherwise; they are not simultaneous confidence bands.

CliffWalking. TIS requires $.65$–$.84$ of learned-occupancy queries, $.32$–$.45$ of rollout queries, and $.84$–1.05 of learned-mean queries. Against learned mean, empirical envelopes are below one at the 12 strictest targets and above one at none; uniform's best tested RMSE exceeds TIS's worst, censoring that comparison in TIS's favor. *Inventory.* Ratios are $.55$–$.72$ versus learned occupancy, $.58$–$.79$ versus learned mean, $.26$–$.30$ versus rollouts, and $.50$–$.83$ versus uniform, with envelopes below one at all 21 targets; only the uniform comparison uses the empirical bootstrap. *FinQA.* On the 25–800-query grid, the anchor uses $.10$–$.17$ of uniform's charged queries on Qwen and $.23$–$.52$ on Phi, with empirical envelopes favoring it at all 21 targets. For Phi ordinary review, envelopes favor the anchor at 14 targets versus occupancy and the 10 strictest versus rollouts, while the baselines win at 4 and 8 targets; for unit-check review these counts are 12 and 6 for the anchor, and 0 and 12 for the baselines. Other contrasts are unresolved, and on Qwen both strong baselines require fewer queries at every target.

A single target can be misleading: uniform's largest-budget RMSE cannot rank sample-only CliffWalking designs because all reach it at the smallest tested budget; inventory shows savings

Table 12: Risk sensitivity at $\alpha = .2$, 400 queries/kernel: TIS MSE/uniform MSE with uniform-minus-TIS z (positive favors TIS); GLM $H = 4$ is unresolved.

Generator	$H = 2$	$H = 4$	$H = 6$
Qwen3-4B	.195 (18.8)	.290 (14.4)	.613 (6.2)
Phi-4-mini	.115 (37.4)	.219 (25.0)	.521 (8.0)
Granite-4.2-8B	.120 (43.8)	.084 (42.6)	.088 (37.1)
Mistral-24B†	.068 (49.5)	.175 (31.5)	.289 (16.9)
Qwen3-32B†	.170 (33.5)	.232 (27.5)	.368 (16.2)
GLM-4-32B†	.345 (20.6)	.929 (1.1)	**1.52** (-6.2)

across the studied range, but its half-budget crossing remains unresolved. **Computation is separate.** Table 13 times complete sampling-and-estimation pipelines on an Intel i9-13900H with one BLAS/OMP thread (ten replications after two warm-ups). Uniform has no pilot; occupancy needs only the forward-visitation solve; TIS, the anchor, learned mean, and the occupancy–mean blend include the influence solve, with the latter three matching TIS within 4%, so the tail-specific score adds no computation over the strongest mean-based baseline. FinQA timing covers all 50 Phi ordinary-review questions (11–197 grid atoms). These frozen-law fixed-budget times are not live-call costs or time-to-equal-accuracy.

Table 13: CPU milliseconds per replication for the full sampling-and-estimation pipeline at fixed query budgets; FinQA is the 50-question Phi ordinary-review panel. These are frozen-law replay times, not live-call costs.

Task	Budget	Uniform	Occ.	Mean	TIS	Anch.	Mean blend	Rollout
CliffWalking	400	110	122	394	394	396	390	**32**
Inventory	1,200	**5.0**	6.5	10.7	10.9	10.8	11.2	5.2
FinQA panel (50 q.)	200	658	823	2,031	2,015	2,039	2,040	**298**

E.3 FinQA terminal-risk protocol and full results

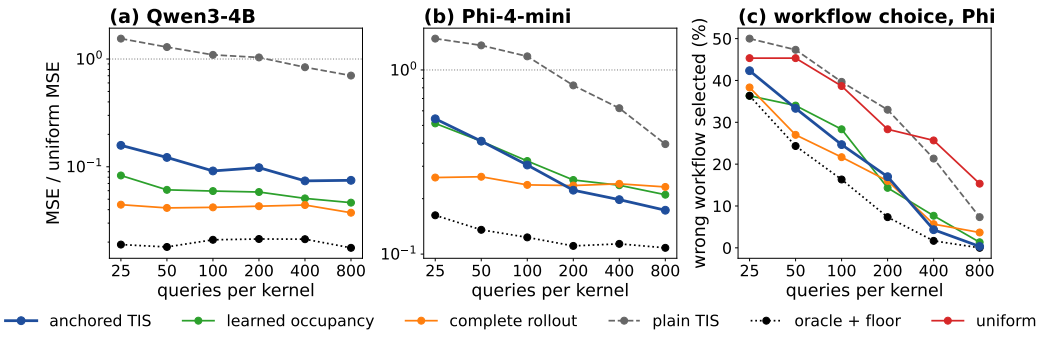


Figure 4: **FinQA: the anchor beats occupancy and rollouts on Phi at 400/800 queries; Qwen favors the baselines.** (a,b) Mean panel-MSE/uniform-MSE ratio across workflows; (c) Phi wrong-selection rate. Table 14 gives per-workflow results.

This held-out study supports **Q4**: does the frozen anchor improve estimation of numerical failure severity and workflow selection (Figure 4)?

Source and screening. FinQA development items (Chen et al., 2021) require finite nonzero executable gold answers and fit the context budget. Each eight-candidate bank is generated before gold-based scoring and admitted only if severity-utility spread is at least .25; this screen was fixed after 16 of the first 20 banks had constant utility, so the study targets material severity disagreement. Development admitted 20/170 items; after protocol freeze, a disjoint held-out scan admitted 50/312.

Workflow and kernels. Candidate banks are generated at temperature .8 and frozen, with parsing failures kept as invalid candidates. The root action is `select`; reviews use the MMLU confidence bands (Table 8). Prompts contain the financial context, question, full bank, current candidate–confidence pair, and review instruction, but no history or stage index. Two constrained output tokens determine the next state. The root plus 72 pairs gives 73 query groups per question, generator, and workflow, shared across stages. Ordinary and unit-and-sign workflows share the root prompt and policy but use different review templates, hence different transition laws. Runs have one selection and two reviews ($H = 3$), with $\alpha = .1$.

Severity utility. For gold y^* and candidate y, let $e_0 = \min(|y - y^*|, |y/100 - y^*|, |100y - y^*|)$, $t = .005|y^*|$, and $e = \max(0, e_0 - t)$. Terminal utility is $u = 1 - e/(e + |y^*|) \in [0, 1]$, with $u = 0$ for unparseable candidates. Binary correctness would collapse magnitude: for $R \in \{0, 1\}$ with failure probability p_{fail}, lower CVaR is $\max\{\alpha - p_{\text{fail}}, 0\}/\alpha$. A post-hoc unit-aware rerun changes screening for 3 development and 5 held-out questions; at 200 queries anchor/uniform MSE is $.096/.100$ on Qwen and $.226/.186$ on Phi for ordinary/unit-check review. The anchor remains resolved against uniform, trails occupancy/rollouts on Qwen, and beats occupancy on Phi; Phi's decision gap widens from 6.7×10^{-4} to 5.6×10^{-3}. The frozen metric defines the primary results.

Reward and exact target. Let $u(s)$ be candidate severity utility, independent of confidence, and $u(\varnothing) = 0$. For sampled next state S',

$$R(s, S') = \frac{u(S') - u(s) + 1}{2} \in [0, 1], \qquad G_H = \sum_{t=0}^{H-1} R(S_t, S_{t+1}) = \frac{H + u(S_H)}{2}. \tag{67}$$

Intermediate utilities telescope, so a temporary improvement that is later reversed does not improve the return. The h-step return-to-go support is $\{(u_j - u_i + h)/2\}$; the grid contains these values and the endpoints, at most $H(C+1)C+2$ atoms ($C = 8$; at most 197 observed). The categorical recursion is exact and matches exact-law CVaR to machine precision, with $C_{\text{ret}} = [H + \text{CVaR}_\alpha(u(S_H))]/2$. All reported CVaRs, gaps, regrets, and MSEs use this scale; converting back by $2\widehat{C}_{\text{ret}} - H$ multiplies MSE by four and doubles gaps without changing within-setting ratios or rankings.

Freeze and audits. Development used 20 screened questions and 100 replications to check parsing (142/160 candidates valid), the .25 spread screen, positive margins, oracle headroom (Qwen oracle+floor .006–.013 of uniform MSE at 100–200 queries), and replication noise; the latter informed held-out size without guaranteeing power. Before calibration, the protocol fixed 50 questions, 300 replications, generators, workflows, anchor, floor, and budgets 25–800 queries per kernel. All four held-out calibrations (3,650 kernels each) pass provenance and generation audits. All 50 margins are positive, but total influence is zero for 3 Qwen questions per workflow and 13/14 Phi questions (ordinary/unit-check), so every allocation has zero leading variance there although higher-order error can remain.

Table 14 gives the complete frozen-metric results through 800 queries; the post-hoc unit-aware sensitivity is summarized above. At 100–200 queries all eight anchor/uniform contrasts are resolved ($z = 8.1$–30.3), while plain TIS is unresolved in seven. At 400/800 queries the anchor beats both strong baselines in both Phi workflows: MSE is .81–.86 of occupancy ($z = -8.4$ to -4.8) and .68–.88 of rollouts ($z = -8.4$ to -3.1), using z for anchor minus baseline. On Qwen, occupancy and rollouts remain better. All contrasts use shared conditional-query streams, a separate rollout stream, and 300 replications.

Workflow decision and coupling sensitivity. Table 15 reports 25–200-query selection rates. Phi's ordinary/unit-check return-CVaR gap is only 6.7×10^{-4}; at 100/200 queries, discordant-pair tests resolve the anchor over uniform ($z = 4.2, 3.5$) but not over occupancy or rollouts ($|z| \leq 1.7$), and the budget-200 regret reduction is 7.6×10^{-5}. Workflow templates share common random numbers, which reduce comparison noise without changing either workflow's marginal MSE. A retrospective re-pairing check shows that Qwen's exceptionally low paired selection errors partly reflect this cancellation; because re-pairing is not fresh simulation, the frozen-protocol rates remain primary.

A separately committed skeptical-template follow-up on Phi had a much larger workflow gap and made the strong methods essentially error-free at small budgets, so it did not distinguish the anchor from occupancy or rollouts. This reinforces why the near-tie frozen study, rather than the easier follow-up, is the informative workflow-selection test.

Table 14: FinQA frozen-metric estimation: uniform panel MSE, anchor-to-baseline MSE ratios, and paired uniform-minus-anchor z over 300 replications.

Gen.	Workflow	Budget	Unif. MSE	**Anch./unif.**	Anch./occ.	Anch./roll.	z
qwen	ordinary	25	$6.2\,e{-5}$	0.152	1.84	3.52	+5.3
qwen	ordinary	50	$2.8\,e{-5}$	0.127	2.10	2.81	+6.5
qwen	ordinary	100	$1.6\,e{-5}$	0.089	1.47	2.07	+8.1
qwen	ordinary	200	$6.7\,e{-6}$	0.097	1.77	2.30	+9.9
qwen	ordinary	400	$3.3\,e{-6}$	0.075	1.41	1.62	+10.0
qwen	ordinary	800	$2.0\,e{-6}$	0.077	1.66	2.11	+10.2
qwen	unit check	25	$1.1\,e{-4}$	0.165	1.97	3.59	+7.7
qwen	unit check	50	$5.7\,e{-5}$	0.117	1.90	3.10	+10.2
qwen	unit check	100	$2.8\,e{-5}$	0.094	1.60	2.28	+11.5
qwen	unit check	200	$1.2\,e{-5}$	0.100	1.61	2.26	+14.3
qwen	unit check	400	$6.9\,e{-6}$	0.073	1.50	1.74	+15.6
qwen	unit check	800	$3.6\,e{-6}$	0.073	1.56	1.89	+15.5
phi	ordinary	25	$3.9\,e{-4}$	0.540	1.10	1.96	+16.7
phi	ordinary	50	$2.0\,e{-4}$	0.431	1.03	1.49	+21.4
phi	ordinary	100	$1.2\,e{-4}$	0.282	0.93	1.10	+27.2
phi	ordinary	200	$5.9\,e{-5}$	0.216	**0.88**	**0.84**	+30.3
phi	ordinary	400	$3.1\,e{-5}$	0.198	**0.86**	**0.77**	+31.5
phi	ordinary	800	$1.5\,e{-5}$	0.184	**0.82**	**0.68**	+27.1
phi	unit check	25	$4.5\,e{-4}$	0.546	1.02	2.23	+15.7
phi	unit check	50	$2.5\,e{-4}$	0.390	0.97	1.65	+21.0
phi	unit check	100	$1.3\,e{-4}$	0.327	0.97	1.50	+23.3
phi	unit check	200	$7.3\,e{-5}$	0.229	0.88	1.07	+26.9
phi	unit check	400	$3.5\,e{-5}$	0.197	**0.81**	**0.88**	+28.9
phi	unit check	800	$1.9\,e{-5}$	0.162	**0.83**	**0.84**	+26.3

Table 15: FinQA frozen-metric workflow selection at 25–200 queries/kernel: exact panel CVaRs, their gap, and wrong-selection percentages under shared workflow streams. Zero means no errors in 300 replications.

Gen.	Budget	C_{ord}	C_{unit}	Gap	Unif.	TIS	Occ.	Roll.	**Anch.**	Oracle +floor
qwen	25	1.87304	1.87262	+0.00041	**0.0**	2.0	**0.0**	**0.0**	0.7	0.0
qwen	50	1.87304	1.87262	+0.00041	**0.0**	0.3	**0.0**	**0.0**	**0.0**	0.0
qwen	100	1.87304	1.87262	+0.00041	**0.0**	0.3	**0.0**	**0.0**	**0.0**	0.0
qwen	200	1.87304	1.87262	+0.00041	**0.0**	1.0	**0.0**	**0.0**	**0.0**	0.0
phi	25	1.76764	1.76697	+0.00067	45.3	50.0	**36.3**	38.3	42.3	36.3
phi	50	1.76764	1.76697	+0.00067	45.3	47.3	34.0	**27.0**	33.3	24.3
phi	100	1.76764	1.76697	+0.00067	38.7	39.7	28.3	**21.7**	24.7	16.3
phi	200	1.76764	1.76697	+0.00067	28.3	33.0	**14.3**	16.0	17.0	7.3

E.4 Inventory disruption family

This prespecified breadth family supports **Q2** by varying three axes of the seasonal inventory benchmark: base disruption probability $\{.01, .04, .12\}$ (plus .02 at capacity), disruption loss of $1, 2$, or 3 remaining units, and two fixed policies (the original seasonal base-stock targets or those targets plus one unit, capped at capacity). Rewards, $\{0, .5, 1\}$ quantization, demand laws, $H = 8$, and the grid are unchanged, producing 18 cases with 41 or 48 retained blocks. Before simulation, exact enumeration confirmed distinct kernel laws in all 18 cases, positive categorical margins (including two near .001), CVaR .675–1.260, and V_{unif}/V^* from 1.49 to 2.28. Budgets are 150, 600, and 1,200 queries per block with 300 replications; methods, pilot, floor, and rounding match Section E.2.3. Absolute RMSE tolerances $.02, .01, .005$ were declared with the family.

Table 16: Inventory family: resolved lower-MSE cases (of 18) at 1,200 queries/block, using paired $|z| \geq 2$; none resolve in the comparator's favor.

Method	occupancy	learned mean	occ.+uniform	occ.+mean	rollouts
TIS	18	17	18	18	18
anchored TIS	18	17	18	17	18

Across the 18 cases, median MSE/uniform at 150/600/1,200 queries per block is $.74/.62/.61$ for TIS, $.82/.69/.69$ for the anchor, $1.05/.92/.93$ for learned occupancy, $.98/.86/.84$ for learned

mean, $1.11/.99/.95$ for the occupancy–uniform blend, $1.03/.87/.86$ for the occupancy–mean blend, $2.72/2.45/2.56$ for rollouts, and $.56/.51/.53$ for oracle+floor. At RMSE .02, TIS first reaches the target by 150 queries in 2 cases and by 600 in 15; uniform, occupancy, and the uniform blend require 1,200 in 6. At .01, TIS reaches 11 cases, uniform/occupancy 8, and rollouts none; at .005, only TIS reaches the target (2 cases). Plain TIS leads the anchor throughout, consistent with anchoring acting as insurance rather than a gain when pilots are reliable.

E.5 Blending controls

To isolate **Q5**, that is whether the tail score itself drives the anchor, we compare two equally regularized controls. With floored component shares $w = (1-\lambda)\hat{p} + \lambda u$, they use $\frac{1}{2}w_{\mathrm{occ}} + \frac{1}{2}u$ and $\frac{1}{2}w_{\mathrm{occ}} + \frac{1}{2}w_{\mathrm{mean}}$, with no second floor; hence all three blends differ only in the component mixed with occupancy. The uniform blend equals occupancy with floor $(1+\lambda)/2$ and coincides with the anchor when the tail pilot falls back to uniform. All blends pay the same pilot and share coupled query streams. On held-out FinQA (2 generators $\times$ 2 workflows $\times$ budgets 100–800; 300 replications), the anchor is resolved better than the uniform blend in all 16 cells (MSE ratios .62–.97, decreasing with budget), while the mean blend has lower point MSE in all 16 and is resolved better in 15 (anchor/mean ratios 1.17–1.54 on Qwen, 1.03–1.10 on Phi). Learned mean alone can reach $1.4\times$ uniform, but occupancy–mean blending is competitive. Under MMLU-Pro confident-error utility (6 generators, $H = 6$, $\alpha = .1$, 100–800 queries, 300 replications), the anchor beats the uniform blend and learned occupancy in all 24 cells and the mean blend in 23 (ratios .76–.97; Phi-4-mini at 100 is unresolved). Rollouts are better for five generators at 100 queries, but the anchor is better for all six at 800 (ratios .55–.89); plain TIS has 1.4–$26\times$ the mean blend's MSE. Under Brier utility (100–400 queries), the anchor beats the uniform blend and occupancy in all 18 cells, the mean blend in 11, is worse in none, and is unresolved in 7 (all Phi-4-mini and Qwen3-32B budgets, plus Qwen3-4B at 100). On the high-stakes panel (Qwen3-4B, Phi-4-mini, Qwen3-32B), it beats the uniform blend and occupancy in all 9 cells and the mean blend in 5; the mean blend wins 2 (Phi-4-mini at 200/400). Brier comparisons with rollouts again cross over with budget: rollouts lead at small budgets, while the anchor leads four of six generators at 400.

E.6 Rare failures and the tail–mean coincidence

Proof of Proposition 2. Let F^{-1} be the quantile function of $X = G_H(s_0)$. By assumption, $X \le x^\star$ almost surely and $\Pr(X < x^\star) = \pi < \alpha$. Hence $F^{-1}(u) = x^\star$ for every $u \in (\pi, 1)$, and boundedness gives

$$\begin{aligned}\alpha\ \mathrm{CVaR}_\alpha(X) &= \int_0^\alpha F^{-1}(u)\,du \\ &= \int_0^1 F^{-1}(u)\,du - \int_\alpha^1 F^{-1}(u)\,du \\ &= \mathbb{E}[X] - (1-\alpha)x^\star.\end{aligned}$$

This proves the stated identity.

By exactness, the grid contains every attainable partial return generated by the fixed declared outcome spaces. Hence changing only the kernel laws changes probabilities but not the structural maximum $x^\star$, and the categorical root law continues to equal the true return law. In a finite horizon, the root return law depends continuously in total variation on the finite collection of group laws; therefore $\Pr(X < x^\star)$ remains below α throughout a sufficiently small neighborhood because the baseline gap $\alpha - \pi$ is strictly positive. On this neighborhood,

$$C_{\alpha,K} = \mathrm{CVaR}_\alpha(X) = \frac{1}{\alpha}\mathbb{E}[X] - \frac{1-\alpha}{\alpha}x^\star.$$

Let $J(P) := \mathbb{E}_P[X]$. Write $v_h(s) = \mathbb{E}[G_h(s)]$ and let $\mu_h(s)$ be the probability, under the fixed policy and population laws, of visiting state s with h steps remaining. Differentiating the ordinary mean Bellman recursion shows that, for a shared stationary group $g = (s, a)$, a centered groupwise influence function for J is

$$\psi_{s,a}(W) = \sum_{h=1}^{H} \mu_h(s)\pi_h(a \mid s)\Big\{R + v_{h-1}(S') - \mathbb{E}_{P_{s,a}}[R + v_{h-1}(S')]\Big\},$$

with the analogous single-row expression in the untied model. This is exactly the ordinary mean-return influence used by the learned-mean design in Appendix E.2. Consequently, for every DQM direction $s = (s_g)$, $\dot{J}_s = \sum_g \mathbb{E}_{P_g}[\psi_g s_g]$. Differentiating the affine identity above and using Equation 38 gives

$$\sum_g \mathbb{E}_{P_g}[\phi_g s_g] = \dot{C}_s = \frac{1}{\alpha}\dot{J}_s = \frac{1}{\alpha}\sum_g \mathbb{E}_{P_g}[\psi_g s_g].$$

Both ϕ_g and ψ_g are centered. Because the product tangent space contains an arbitrary $L_0^2(P_g)$ direction in each group, equality for every score direction implies $\phi_g = \psi_g/\alpha$ in $L_0^2(P_g)$ for each group. Consequently the tail and mean influence standard deviations satisfy $\sigma_g = \mathrm{sd}(\psi_g)/\alpha$. If these scales are not all zero, normalizing them gives identical Neyman shares; if they are all zero, both objectives have zero first-order variance for every allocation. This proves the allocation claim. □

This diagnostic supports **Q5** by testing when tail-specific allocation should differ from mean allocation. All experiments use closed grids exact on their declared supports. Normalize each nonzero tail- or mean-influence scale vector to sum one, representing an all-zero vector by uniform; then $\frac{1}{2}\sum_g |p_g - m_g| = 0$ whenever Proposition 2 applies. The median distance is .000 on held-out FinQA for both generators and workflows, with exact zero on 45–51% of questions. On 25-question MMLU-Pro samples at $H = 6$, $\alpha = .1$, Brier-utility medians are .24 (Qwen3-4B), .09 (Phi-4-mini), .29 (Granite-4.2-8B), .30 (Mistral-24B), .16 (Qwen3-32B), and .27 (GLM-4-32B); the high-stakes panel gives .33, .08, .18 for Qwen3-4B, Phi-4-mini, Qwen3-32B, and confident-error utility gives .32, .18, .32 for Qwen3-4B, Phi-4-mini, GLM-4-32B. Granite, Mistral, Qwen3-32B, and the high-stakes Phi/Qwen3-32B distances were computed after the blending runs.

Prospective divergence test. We therefore fixed a rule before simulating five new settings: median distance $\geq .24$ predicts that the anchor has resolved lower MSE than the mean blend in at least two of three budgets (100/200/400 queries) and higher MSE in none; distance $\leq .18$ predicts at most one resolved win; intermediate values make no prediction. The five settings were high-stakes confident-error Qwen3-4B (.318, win), Phi-4-mini (.169, no advantage), Qwen3-32B (.216, none), and cautious-policy Qwen3-4B with Brier (.195, none) or confident-error (.269, win). All use closed grids, $H = 6$, $\alpha = .1$, and 300 replications. Against the mean blend, MSE ratios (paired z) at 100/200/400 are $.88/.84/.83$ $(-6.2/-8.5/-9.7)$ for high-stakes Qwen3-4B, $1.03/1.02/.99$ $(+2.6/+1.4/-0.6)$ for high-stakes Phi, and $.96/.95/.92$ $(-3.0/-4.0/-5.3)$ for cautious confident-error Qwen, so all three predictions hold. The two unpredicted settings give $.97/.94/.91$ $(-1.3/-4.3/-6.0)$ for high-stakes Qwen3-32B and $.96/.92/.88$ $(-3.3/-5.5/-8.7)$ for cautious Brier Qwen. In all 15 cells the anchor also beats learned occupancy and the uniform blend; at 400 queries it beats complete rollouts in four settings (ratios .74–.85) and ties the fifth (.99).

The confident-error utility scores a correct response with confidence c as $(1 + c)/2$ and a wrong response as $(1-c)^2/2$, making confident errors nearly worthless; it preserves the Brier per-response ordering but has a much heavier lower tail.

E.7 FinQA with calculator faults

This robustness check extends **Q5** to exogenous tool errors while preserving the FinQA terminal-severity objective. Each review prompt includes an automated calculator report for the current candidate; independently after each model call, the report is correct with probability $1-p$ and multiplied by 100 with probability p. The root, 72 correct-report states, and 72 faulted-report states define 145 queryable prompt laws per question; an outcome is the model response together with the fault indicator, so one query still costs one model call. The laws were calibrated exactly for both generators on the 50 held-out questions, and the same design was declared for ordinary and unit-check review.

The perturbation is informative because influence and visitation react differently. For Qwen, faulted states carry median tail/mean/occupancy shares $5.8/6.2/0.7\%$ at $p = .01$ and $38/42/6.7\%$ at $p = .10$; for Phi the corresponding shares are $1.0/0.9/0.7\%$ and $9.8/9.0/6.7\%$; 12/50 Phi questions have no first-order tail signal. Thus tail and mean influence remain close even when occupancy can be very different. Across 100–400 queries per kernel, anchor/uniform MSE is .034–.050 for Qwen and .081–.144 for Phi in ordinary review, with similar .041–.050 and .069–.144 ranges under unit check. Yet rollouts are usually strongest, the mean blend beats the anchor throughout Qwen, and the anchor

trails occupancy on Qwen while beating it on Phi. Plain TIS can have 7–29× the mean blend's MSE. Qwen's oracle remains only .007–.010 of uniform MSE, locating the gap in pilot learning rather than the population influence signal; several Phi questions instead have near-zero margins. Median tail–mean distance is .000 for Qwen and .03–.06 for Phi, well below the no-advantage regime identified in Appendix E.6. The tool-fault study therefore supports the diagnostic's negative prediction: a tail score can be highly informative in population yet unnecessary relative to a smoother mean score when their normalized allocations nearly coincide.

E.8 Longer review loops

This extension supports the main-text longer-loop claim and re-tests **Q4–Q5**: because prompts omit stage index, the frozen kernels define longer loops without new model calls. Before simulation we declared MMLU-Pro confident-error runs at $H = 8, 10$ for Qwen3-4B, Phi-4-mini, and GLM-4-32B, plus held-out FinQA runs at $H = 6$ for both generators and workflows (100–400 queries per kernel, 300 replications), using the $H = 6$ MMLU and $H = 3$ FinQA blending runs as references and recording the divergence values and three predictions.

Divergence rule. Appendix E.6's rule holds in 6/8 settings it decides: GLM (distance .36/.39) beats the mean blend at every budget at both horizons; all four FinQA settings (distance .000–.098) show no advantage; Qwen3-4B (distance .26 at both horizons) has anchor/mean ratios .92–.98 but resolves only at $H = 8$, 400 queries, so its two win predictions fail. Phi-4-mini has distance .19 (no prediction) and wins 5/6 cells.

Kernel reuse. Table 17 shows that longer shared-kernel loops increasingly favor the anchor for MMLU and FinQA Phi but not near-deterministic FinQA Qwen; the prediction holds in 5/7 settings. For Phi at $H = 6$, anchor MSE is .29–.64 of rollout MSE at every tested budget in both workflows, i.e. 1.6–3.4× lower and 2.4–3.4× lower at 400 queries, which are the reductions quoted in the abstract.

Table 17: **Longer review loops favor the anchor over rollouts.** Anchored-TIS/rollout MSE ratio (paired z; negative favors the anchor), 300 replications. The Phi $H = 6$, 400-query ratios .29/.41 give the 3.4 × /2.4× reductions in the abstract.

	FinQA $H = 3$			FinQA $H = 6$		
Generator, workflow	100	200	400	100	200	400
Phi-4-mini, ordinary	1.10 (+1.9)	.84 (−4.4)	.77 (−5.2)	.36 (−22.1)	.30 (−22.5)	.29 (−22.4)
Phi-4-mini, unit check	1.50 (+10.0)	1.07 (+1.8)	.88 (−3.1)	.64 (−10.4)	.47 (−17.3)	.41 (−19.7)
Qwen3-4B, ordinary	2.07 (+6.1)	2.30 (+5.9)	1.62 (+3.4)	2.64 (+6.5)	2.03 (+5.1)	2.02 (+5.3)
Qwen3-4B, unit check	2.28 (+9.5)	2.26 (+8.6)	1.74 (+5.4)	2.73 (+8.8)	2.09 (+7.2)	2.13 (+7.3)
MMLU-Pro, 400 queries	$H = 6$		$H = 8$		$H = 10$	
Qwen3-4B	.80 (−6.6)		.65 (−7.5)		.67 (−5.5)	
Phi-4-mini	.94 (−1.7)		.86 (−3.5)		.84 (−4.0)	
GLM-4-32B	.90 (−2.8)		.95 (−1.1)		.89 (−2.8)	

Pilot risk. From $H = 6$ to 10, the plain-TIS/anchor MSE ratio at 400 queries changes 7.4 → 7.6 for Phi-4-mini, 30.1 → 32.0 for GLM, and 19.0 → 14.4 for Qwen3-4B, so this prediction holds in 2/3 settings. Across all 18 MMLU-Pro cells, the anchor attains .07–.19 of uniform MSE and is resolved better than learned occupancy and the uniform blend in every cell.

E.9 Selecting the allocation from the pilot

This section operationalizes **Q5**: the population tail–mean distance is unavailable, but every learned design already draws a uniform pilot. Before evaluation we declared: for each setting and budget, compute the replication-0 pilot distance between normalized tail and mean influence shares for each question, take the median, use the anchor if it is at least .21 (the midpoint of the population thresholds), and otherwise use the occupancy–mean blend. A choice is wrong only if the selected design has resolved higher MSE ($|z| \geq 2$) than the alternative.

Across 46 blending settings (148 setting–budget cells), the rule is wrong in 7; selected-design MSE relative to the better design has median 1.00. Pilot and population medians have Spearman correlation .86, and replication-0 decisions agree with replications 1–4 in 91% of cases. Three errors are Phi-4-mini confident-error cells with pilot medians .197–.206 just below threshold, costing 3–6% MSE. Four are FinQA Qwen3-4B $H = 6$ cells where near-determinism makes a small pilot overstate distance (pilot .27–.57, population .000), giving the anchor 1.4–2.3× mean-blend MSE. Thus the statistic is reliable except when failures are too rare for the pilot to observe, precisely the regime in which Proposition 2 removes the need for tail-specific allocation.

E.10 Mechanism: what the pilot sees and what anchoring changes

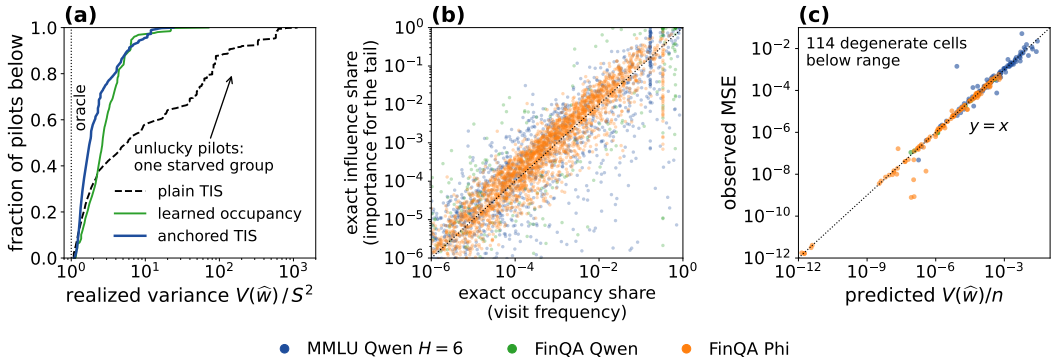


Figure 5: **Anchoring reduces the variance penalty from poor pilots.** (a) Realized/oracle leading variance; (b) occupancy versus tail-influence shares; (c) observed MSE versus the no-fit prediction $V(w)/n$. Rollouts are excluded.

Figure 5 explains **Q3**'s failure mode at the query-share level. Equation 51 penalizes underallocation by dividing squared allocation error by assigned weight. We replay the recorded pilots, floors, and rounding and use population scales to diagnose $V(\widehat{w}) = \sum_g \sigma_g^2/\widehat{w}_g$ for MMLU Qwen ($H = 6$, 400 queries), FinQA Qwen ordinary review (200), and FinQA Phi ordinary review (200). All questions enter; normalization by $S^2 = (\sum_g \sigma_g)^2$ excludes their 1, 3, and 13 zero-influence questions. Rollouts are excluded because they estimate under a different sampling functional.

Median correlations between population influence and occupancy shares are $.74, .85, .62$ in the three settings, and the bottom-occupancy half of kernels carries essentially no tail influence (Figure 5(b)), unlike Proposition 1's equal-visitation construction. For MMLU pilots, median/90th-percentile/maximum $V(\widehat{w})/S^2$ is $5.6/136/1{,}137$ for plain TIS, $1.8/5.8/24$ for the anchor, and $2.6/6.2/71$ for occupancy; a single starved group can dominate variance (90th-percentile top-contribution share 1.00). The anchor multiplies the most-starved group's weight by median factors $7.5, 20, 4.9$ across the three settings (MMLU 90th percentile 99). On near-deterministic FinQA Qwen, pilots often see point masses, estimated scales vanish, and TIS approaches uniform: median V/S^2 is 72.4 with 73 groups versus 3.3 for the anchor. These are leading-variance diagnostics, not finite-sample MSE guarantees.

Using actual main-sample counts, pilot-averaged $V(\widehat{w})/n$ predicts MSE without fitted constants in the examined nondegenerate cells: median $\log_{10}$(observed/predicted) is $-.001$ over 98 MMLU question–method pairs and $-.006$ over 135 Phi pairs, with 10th–90th percentiles within $\pm.19$ (Figure 5(c)). For near-deterministic Qwen questions, 300 replications can miss rare deviations carrying much of the variance, so observed MSE can be orders of magnitude lower; the plot flags 114 below-range FinQA pairs (96 Qwen, 18 Phi). Agreement elsewhere is a retrospective check of the variance formula, not a finite-budget guarantee.

F Additional Related Work

Table 18 summarizes the closest foundations; the connections below extend Section 6.

Table 18: Established foundations and the additional results for the conditional-query CVaR problem. The allocation rule is Neyman allocation; the work here identifies and learns the Bellman influence scales it needs.

Established starting point	Additional result here
Return-law and functional inference under specified sampling (Zhang et al., 2025); quantile-based efficiency (Cheng et al., 2026).	An explicit categorical-CVaR influence for each conditional kernel, including its joint reuse across stages. Theorems 1 and 2 give the variance as a function of query shares and its fixed-design efficiency interpretation.
Neyman allocation and learning unknown stratum variances (Neyman, 1934; Étoré & Jourdain, 2010; Carpentier et al., 2015).	The influence function depends on the unknown model and quantile. Theorem 3 controls learning those quantities, quantile errors, and Bellman remainders to obtain oracle variance and normalized MSE, including pilot cost and individual zero-influence groups.
Trajectory collection for mean policy evaluation (Mukherjee et al., 2022; 2024).	Independent conditional queries for a fixed policy's tail functional. Proposition 1 shows why visitation and mean-optimal allocation can miss the relevant signal. Comparisons with complete rollouts are empirical; the product-model efficiency theorem does not cover their different sampling experiment.

The distinction from adjacent work is mainly the design variable. Distributional and quantile-based RL develop inference or efficiency under specified data laws (Zhang et al., 2025; Cheng et al., 2026; Peng et al., 2024; Peng & Zhang, 2026), while generative-access and offline analyses study estimation error under fixed access models (Rowland et al., 2024; Peng et al., 2025; Chandak et al., 2021; Wu et al., 2023; Hong et al., 2025). Stratified and adaptive experimental design learn Neyman allocations when the within-stratum target is already defined (Carpentier et al., 2015; Dai et al., 2023); our score itself depends on an unknown Bellman continuation model and CVaR cutoff, which is why Theorem 3 must control learning the influence function as well as its allocation. Risk-sensitive control and logging-policy design change the policy or trajectory distribution (Bäuerle & Ott, 2011; Zhu et al., 2024; Douglas et al., 2026); here the policy and conditional laws stay fixed and only the independent conditional-query counts change. Active testing allocates effort across benchmark items (Nguyen et al., 2018; Kossen et al., 2021; Maia Polo et al., 2024; Li et al., 2025); our experiments allocate within an item's stochastic workflow, and combining the two levels is a natural extension.