WHICH TEAMS MUST BE EVALUATED? COST AND ROBUSTNESS LIMITS OF COMPOSITIONAL AGENT EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating every combination of frozen agents can be costly. We study simultaneous value estimation with charges for team preparation and each evaluation trial, conditional on a supplied valid bound on mean-model error. Under fixed Hoeffding certification, we identify exactly when sharing observations cannot reduce any team's direct sample requirement. A tensor identity computes this boundary for complete bipartite catalogues. For a four-role path, the cheapest plan shifts from sixteen teams to twelve with unequal sampling, then eight, as preparation prices rise. With model error, preparing eight rather than sixteen teams can require 39 times as many observations. Positive observation prices preserve the boundary but change the preferred allocation. A finite-pool evaluator compares empirical-Bernstein and direct confidence-sequence procedures, charges selection costs, and confirms the selected procedure on independent observations. Controlled studies distinguish structural sharing, support adaptation, and procedure choice. These results characterize acquisition-cost trade-offs under stated model and certificate assumptions; valid confirmation alone does not guarantee cost-optimal selection or deployment savings.

1 INTRODUCTION

Four roles with two frozen policies already yield sixteen teams. Evaluating every team separately ignores what their shared components may reveal about one another. Modular systems make this question relevant (Wu et al., 2024): performance depends on coordination (Cemri et al., 2025; Kim et al., 2026; Zhu et al., 2025), and reliable comparison requires more than point estimates (Rabanser et al., 2026). We seek simultaneous confidence intervals for the entire team catalogue.

Example-efficient benchmarks reduce the test items needed per system (Maia Polo et al., 2024; Wang et al., 2025). We instead study sharing observations across teams under a low-dimensional interaction model. Preparing fewer teams can require more observations per team; model error can erase the saving. Following cost-aware evaluation (Kapoor et al., 2025; 2026), we charge preparation and every consumed observation, including pilots. Our theoretical question is which plans minimize this charge under a supplied valid model-error bound. Obtaining that bound is a separate, charged task. Figure 1 distinguishes model dimension, sampled teams (the support), and direct evaluation.

First, we derive a sharp direct-optimality boundary under fixed Hoeffding certification: at or above it, every team needs at least its direct sample count. Below it, sharing can save observations even when every team is sampled. A tensor identity gives explicit boundaries for complete bipartite catalogues; positive observation prices leave the boundary unchanged.

Second, we characterize the four-role path's continuous preparation-cost frontier, including a nonuniform twelve-source phase between full and compact supports. Model error changes which compact support is feasible: halving the prepared teams can require 39 times as many observations. Minimum support, minimum sampling, and minimum total cost are different objectives.

Third, cost-aware catalogue evaluation (CACE) selects a sampling plan; its engine-aware extension (CACE-E) also selects the confidence procedure. It pays for an optional pilot and confirms the choice on independent observations. Controlled studies separate the benefits of sharing, support adaptation,

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

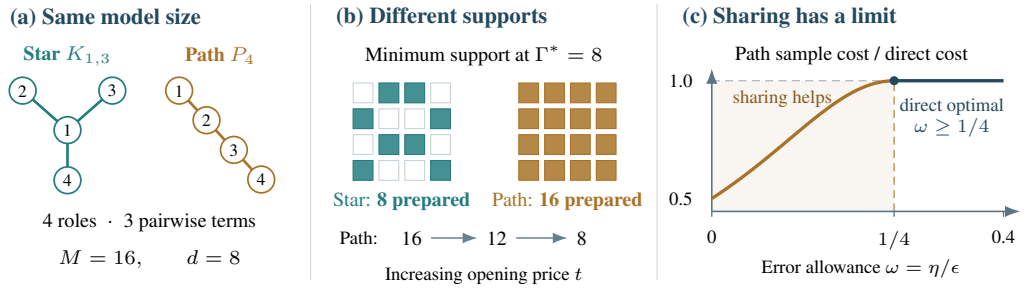

Figure 1: **Structural geometry and evaluation cost.** (a) Equal-dimensional star/path models. (b) Minimum supports attaining the optimal prediction factor (filled cells) and the opening-cost trade-off. (c) Exact zero-setup path ratio, reaching direct sampling at $\omega = 1/4$. Analytic examples within the stated certificate class.

and engine choice. These are conditional acquisition-cost results; deployment savings also require affordable assurance and realistic prices.

2 Evaluation Target and Assurance

A configuration $z \in \{-1, +1\}^h$ selects one of two frozen policies per role. Its value is $V(z) = \mathbb{E}_P Y(z, \xi)$ under a fixed task law, with independent scalar observations $Y \in [0, 1]$. A catalogue $\mathcal{T}$ contains the M target teams whose values are required. Sources are teams available for sampling, and a support is the subset actually sampled. A declared feature matrix X, whose target row is $x_x^\top$, spans a model $\mathcal{Q}$ containing constants. We assume

$$V = X\beta + r, \qquad \|r\|_\infty \leq \eta. \tag{1}$$

The allowance η bounds mean-model error over the source–target domain. A pairwise outcome graph G gives $\mathcal{Q}_G = \text{span}\{1, z_i, z_i z_j : ij \in E_G\}$: edges specify score interactions, not a communication protocol. *Assurance* justifies η independently of confirmation; a fitted residual is insufficient (Appendix A.2). Statistical assurance uses budget δ_A and inference uses δ_I, with $\delta_A + \delta_I \leq \delta$. For deterministic assurance $\delta_I = \delta$; otherwise use δ_I in the certificate formulas.

At declared observation checkpoints (looks) $t \in \mathcal{L}$, the endpoint is simultaneous coverage and half-width ϵ:

$$\mathbb{P}\{V(x) \in I_{x,t} \; \forall x \in \mathcal{T}, t \in \mathcal{L}\} \geq 1 - \delta, \qquad T_\epsilon = \inf\{t : \max_x |I_{x,t}|/2 \leq \epsilon\}. \tag{2}$$

For $0 < \epsilon, \delta < 1/2$, preparing (opening) s configurations and collecting N observations costs $\kappa s + cN$: κ is a one-time preparation charge and c an observation price. Pilots, assurance acquisition, and failures are charged when incurred. Noncompletion receives a declared cap and is reported separately. The exact frontiers concern catalogue-only sources, continuous counts, homogeneous prices, and the fixed Hoeffding certificate; engine-specific acquisition is distinguished below.

3 When Does Sharing Cease to Help?

3.1 A Sharp Direct-Optimality Boundary

Reproducing weights $X^\top w_x = x_x$ recover the target exactly when the declared mean model is correct. For fixed weights, weighted Hoeffding and a target-wise union bound give

$$R_x = \eta \, \|e_x - w_x\|_1 + \sqrt{A \sum_s w_{xs}^2 / n_s}, \qquad A = \tfrac{1}{2} \log(2M/\delta). \tag{3}$$

The first term protects against model error; the second accounts for sampling uncertainty. Direct weights $w_x = e_x$ cancel the residual exactly. An omitted target instead incurs multiplier $1 + \|w_x\|_1$.

With $L = A/\epsilon^2$, $\omega = \eta/\epsilon$, and $q_s = n_s/L$, the precision constraints are

$$\omega \left\| e_x - w_x \right\|_1 + \sqrt{\sum_s w_{xs}^2/q_s} \leq 1 \quad \text{for every } x. \tag{4}$$

Define the omitted-target reconstruction norm and catalogue threshold by

$$\Lambda_x = \min\{\|v\|_1 : X_{-x}^\top v = x_x\}, \qquad \omega_* = \max_x \frac{1}{1+\Lambda_x}. \tag{5}$$

Set $\Lambda_x = \infty$ if omission destroys estimability, with reciprocal zero. This minimum-ℓ_1 geometry (Bos, 2023; Ettehad & Foucart, 2021) determines a sample-cost boundary, not just a reconstruction error.

Theorem 1 (Sharp direct-optimality boundary)**.** *For a finite catalogue, a model containing constants, catalogue-only sources, homogeneous positive observation prices, and fixed reproducing estimators, direct evaluation is continuously cost-optimal under* (4) *if and only if $\omega \geq \omega_*$. Every feasible design then opens every target and assigns $n_x \geq L$ to each. Continuous and integer optimum costs are*

$$M\kappa + cML, \qquad M\kappa + cM\lceil L \rceil. \tag{6}$$

Below the threshold, a strictly cheaper continuous full-support design exists; integer savings need not exist.

For $a = w_{xx}$, reproduction gives $\|e_x - w_x\|_1 \geq (1+\Lambda_x)|1-a|$. Above the threshold, $q_x < 1$ is infeasible. Below it, a small blend of an omitted-target representation into a direct row reduces that target's count. Appendix C.7 proves both directions.

Corollary 2 (Price-invariance of the direct boundary)**.** *The same threshold holds with configuration-specific opening prices $\kappa_x \geq 0$ and observation prices $c_x > 0$. Above it the continuous optimum is $\sum_x \kappa_x + L\sum_x c_x$; below it some continuous full-support plan costs strictly less.*

The per-target lower bound does not depend on prices, and the converse reduces a positively priced count without changing openings. This corollary does not make the optimal sharing support price-invariant.

For a proposed estimable support S, finite-count feasibility requires and is guaranteed by

$$\eta\{1+\Lambda_x(S)\} < \epsilon \quad (x \notin S), \qquad \Lambda_x(S) = \min\{\|v\|_1 : X_S^\top v = x_x\}. \tag{7}$$

This outcome-free screen separates feasible omission from economical omission (Appendix C.5).

3.2 An information obstruction and tensor composition

The count L is certificate-specific. Omission also has an information-theoretic obstruction in $\mathcal{V}_\eta = (\mathcal{Q} + \eta[-1,1]^M) \cap [0,1]^M$.

Proposition 3 (Omission beyond linear inference)**.** *For finite Λ_x, exact knowledge that $V(s) = 1/2$ for all $s \neq x$ leaves the target interval*

$$[\tfrac{1}{2} - D_x(\eta), \tfrac{1}{2} + D_x(\eta)], \qquad D_x(\eta) = \min\{\tfrac{1}{2}, \eta(1+\Lambda_x)\}. \tag{8}$$

For $\Lambda_x = \infty$, the full interval $[0,1]$ remains possible even at $\eta = 0$. If $\epsilon < D_x(\eta)$, any uniformly valid procedure that never observes this target completes with probability at most 2δ for an indistinguishable pair of admissible laws.

A dual extremal function and cancelling residuals construct identical non-target observations with different target means. This optimal-recovery consequence permits nonlinear inference and adaptive non-target sampling; it does not assert Hoeffding sample lower bounds or an arbitrary-inference impossibility at equality (Appendix C.4).

Theorem 4 (Tensor composition)**.** *For finite factor catalogues $\mathcal{T}_j$ and spaces $\mathcal{Q}_j$ containing constants, with full product sources and targets and product span $\bigotimes_j \mathcal{Q}_j$,*

$$\Lambda_x^\otimes = \min_j \Lambda_{x_j}^{(j)}, \qquad \omega_*^\otimes = \max_j \omega_*^{(j)}. \tag{9}$$

The identities include nonreproducible factors under the infinity convention.

Varying one factor while holding all others at their target coordinates gives the upper bound. For two factors, normalized products of dual witnesses attain the reverse bound (Appendix C.8).

Corollary 5 (Complete-bipartite boundary)**.** *For the full $K_{a,b}$ pairwise model, $m = \max(a, b) \geq 2$ gives*

$$\Lambda_x = \frac{m+1}{m-1}, \qquad \omega_* = \frac{m-1}{2m}. \tag{10}$$

The saturated $K_{1,1}$ model always selects direct in the stated certificate class.

The pairwise bipartite span is the tensor product of two additive blocks. Hence $K_{3,3}$ and $K_{3,5}$ have thresholds $1/3$ and $2/5$ without support enumeration. These are direct-certificate boundaries, not predicted crossovers against every confidence engine.

4 Below the boundary: which support is economical?

Let $\pi_u \geq 0$, $\sum_u \pi_u = 1$, be the fraction of observations allocated to team u. For the d binary character features $\psi(u)$, orthonormal under the uniform catalogue law with $\|\psi(u)\|_2^2 = d$, define the information matrix I and worst-case prediction-variance factor Γ:

$$I(\pi) = \sum_u \pi_u \psi(u)\psi(u)^\top, \qquad \Gamma(\pi) = \max_u \psi(u)^\top I(\pi)^{-1}\psi(u). \tag{11}$$

The classical trace bound gives $\Gamma \geq d$, with equality exactly at $I = I_d$; singular designs have infinite factor. The four-role star and path both have $d = 8$, but attaining this optimum requires respectively eight and sixteen sources. Feature-product moments cause the difference, not rank (Appendix C).

For $t = \kappa/(cL)$, exact-model continuous cost satisfies

$$C_H^* = \inf_\pi \{\kappa |\operatorname{supp}\pi| + cL\Gamma(\pi)\}. \tag{12}$$

Theorem 6 (Nominal path frontier)**.** *For the full four-role path,*

$$C_H^*/(cL) = \min\{8 + 16t,\ 6 + 4\sqrt{2} + 12t,\ 16 + 8t\}. \tag{13}$$

The optimal supports have 16, 12, and 8 points, with transitions $\sqrt{2} - 1/2$ and $5/2 - \sqrt{2}$. For the star, $C_H^*/(cL) = 8 + 8t$.

The twelve-source design is nonuniform. At $t = 1$, its cost is $23.657cL$, 1.43% below the best shared extreme. Exact support-exclusion certificates establish a genuine intermediate phase, rather than a large generic improvement (Appendix C.9).

Theorem 7 (Robust path sampling and preparation)**.** *At zero setup price the minimum normalized count is*

$$H(\omega) = \begin{cases} 16/[1 + (1 - 4\omega)^2], & 0 \leq \omega \leq 1/4, \\ 16, & \omega \geq 1/4. \end{cases} \tag{14}$$

At $\omega = 1/5$, the complete preparation frontier is

$$C_R^*(t)/(cL) = \min\{200/13 + 16t,\ 150 + 100\sqrt{2} + 12t,\ 600 + 8t\}, \tag{15}$$

with transitions $t_A = 875/26 + 25\sqrt{2} \approx 69.009$ and $t_B = 225/2 - 25\sqrt{2} \approx 77.145$.

The nominal eight-source support is infeasible at this allowance; a different eight-source support is feasible. The robust full and compact plans need $200L/13$ and $600L$ observations: halving the openings increases observations by a factor of 39. Robustness therefore changes *which teams to open*, not merely the counts. At $\epsilon = .02$, $\delta = .05$, the twelve-source robust phase requires opening prices equivalent to about 557,376–623,085 observations per source, an assigned high-preparation regime. Figure 2 summarizes the exact boundaries; Appendix C.10 gives the constructions and exclusions. Larger bipartite catalogues admit classical product designs with $d \leq s < 4d$ sources, but do not inherit this path frontier (Dey et al., 2005).

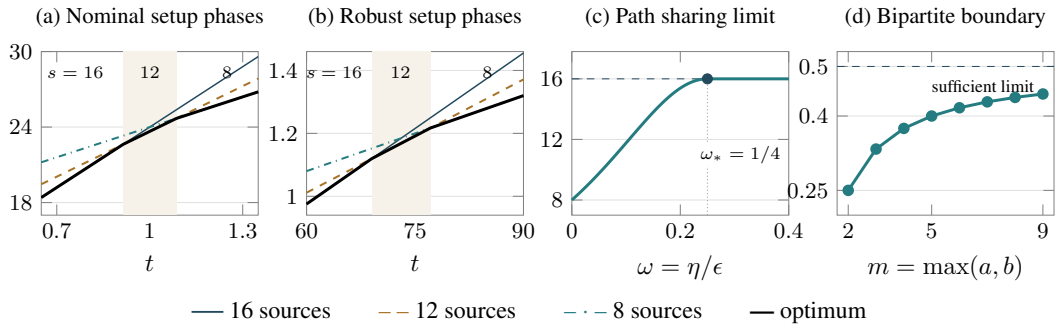

Figure 2: **Preparation and robustness limits under fixed Hoeffding certification.** (a,b) Path transition costs $C/(cL)$ at $\omega = 0$ and $C/(1000cL)$ at $\omega = 1/5$. Shading marks twelve-source phases; black curves are optimal. (c) Zero-setup sampling becomes direct at $\omega = 1/4$. (d) Complete-bipartite thresholds approach $1/2$.

5 Engine-aware catalogue evaluation

A confidence *engine* constructs the intervals and determines when precision is reached. Each candidate fixes an engine, support, reproducing weights, and allocation schedule. The pool includes the solved path designs, classical products, full-support sharing, and direct evaluation (Algorithm 1). The comparisons distinguish the following choices:

Method	**What it chooses or holds fixed**
Direct-EB / Direct-CS	Sample targets separately; use empirical-Bernstein (EB) / confidence-sequence (CS) intervals.
EB-only selector	Choose support, weights, and allocation within EB.
CACE-E	Choose an EB plan or Direct-CS; acquire a pilot only if affordable.
No pilot	Use the same choices as CACE-E, with no selection observations.
Matched EB-only	Use CACE-E's pool, pilot rule, and information, but only EB.

For an empirical-Bernstein (EB) candidate, simultaneous source-variance bounds $\overline{v}_s$ give

$$R_x^{\mathrm{EB}} = \eta \left\| e_x - J w_x \right\|_1 + a_x + \sqrt{2V_x^+ \ell + a_x^2}, \qquad V_x^+ = \sum_s w_{xs}^2 \overline{v}_s / n_s, \tag{16}$$

where $\ell = \log(4M/\delta)$, $a_x = (\ell/3)\max_s |w_{xs}|/n_s$, and J embeds sources into the full domain. Repeated looks use their allocated error budgets. Direct-CS samples the target with the widest interval and uses a gamma-exponential boundary (Howard et al., 2021), also used for its cost forecast. Appendix B.2 specifies this boundary and tuning; Appendix B.5 gives EB variance bounds, conic planning, and rounding (Maurer & Pontil, 2009).

The selection stage starts from a common variance forecast. Let U be the union of eligible candidate supports. A pilot with b observations at each source has charge

$$B_U = \sum_{s \in U} \kappa_s + b \sum_{s \in U} c_s. \tag{17}$$

Acquire it only when $B_U \le \tau \min_p \widehat{C}_p^0$, where $\widehat{C}_p^0$ is the cost forecast before observing outcomes and τ is fixed in advance. This gate prevents pilot openings from erasing a compact plan's advantage. An affordable pilot updates variance forecasts by shrinkage toward a common prior; otherwise use the initial forecast.

After selection has incurred charge C_0, the forecast for candidate p is

$$\widehat{C}_p = C_0 + \sum_{s \in S_p \setminus U} \kappa_s + \sum_{s \in S_p} c_s \widehat{n}_{p,s}. \tag{18}$$

With no pilot, $C_0 = 0$ and $U = \varnothing$. Charge each opening once, including unused pilot sources. Select the cheapest EB candidate only if its forecast is at most $(1-\gamma)\widehat{C}_{\mathrm{CS}}$; otherwise select Direct-CS. The fixed margin γ requires a predicted saving large enough to justify choosing EB.

Freeze the engine, weights, schedule, and tuning before independent confirmation. EB follows its fixed schedule; Direct-CS retains its valid width-adaptive allocation. Pilot observations are charged but excluded from the final estimates.

Proposition 8 (Validity after engine selection)**.** *Let $\mathcal{F}_0$ contain all selection information. Suppose the chosen procedure, conditional on $\mathcal{F}_0$ and valid assurance, has simultaneous failure probability at most δ_I on its confirmation stream. Then arbitrary $\mathcal{F}_0$-measurable selection preserves failure probability at most $\delta_A + \delta_I$.*

Only the selected procedure is confirmed, so no extra union penalty over engines is needed (Appendix B.1). This preserves coverage, not forecast accuracy or global cost optimality. The exact Hoeffding frontiers retain their separate assumptions.

6 EXPERIMENTS

The evaluation separates structural sharing, support selection, and confidence-engine choice by their cost of reaching the requested precision. Each family uses 30 independently trained populations with 50 acquisition repetitions nested within each population. Unless stated otherwise, $\epsilon = .02$ and $\delta = .05$. Cost includes preparation and all consumed observations, including selection pilots excluded from final estimation; capped noncompletion remains in the analysis. Arithmetic mean charged cost is the budget estimand; geometric paired ratios summarize multiplicative comparisons, not expected budget savings. Population-level pairing accounts for reuse (Appendix B.3). Appendices D, E, and B give the task generators, confidence budgets, and reference construction.

6.1 STRUCTURED ACQUISITION AND SUPPORT SELECTION

The controlled scores add node and pair terms for frozen contextual-bandit policies, so the mean model holds by construction. P4/P6 denote four-/six-role paths; B33/B35 denote $K_{3,3}/K_{3,5}$; W6 is a six-role dispatch task. P6 has 32 non-affine targets and permits other source teams. *Target-only sharing* restricts sources to targets; *classical plans* use fixed design constructions. At zero allowance and setup price, sharing saves 26.51%–81.32% against Direct-CS and 11.75%–40.94% against target-only sharing. Selected, classical, and forecast-ranked pool-reference plans coincide: the gains come from sharing, not improved selection (Table 6; protocols and precision curves in Appendix D).

The five operational settings then introduce approximation allowances and opening costs. Four selected EB plans improve on both direct comparators, with savings of 29.13%–60.55% against Direct-EB and 2.95%–28.98% against Direct-CS. High-error B35 selects direct EB, whose arithmetic mean charge is 82.56% above Direct-CS (Table 1). This contrast motivates including the confidence engine in the acquisition decision.

B33 and moderate-error B35 use full support and share estimation. In the B33 control, full-support projection saves 26.06% over the compact design; reweighting raises this to 26.20%, with the increment unresolved. High-setup P4 reduces openings from sixteen to eight but uses 4,110,976 observations versus Direct-EB's 91,954.76. The assigned opening price makes this economical. All selected runs complete. Their reference brackets are $[99.93, 100]\%$ for high-setup P4 and $[100, 100]\%$ otherwise. A bracket spans the confirmed fraction through the fraction not definitely missed, conditional on valid references. A run is confirmed if all reported intervals contain their reference intervals; any disjoint pair is a definite miss, and other cases are indeterminate. These recorded-run classifications are not population coverage confidence intervals (Appendix B).

6.2 ENGINE SELECTION AND THE VALUE OF PILOT INFORMATION

Choosing the confidence procedure avoids the high-error B35 penalty; paid pilots have a conditional additional benefit. CACE-E uses pilot size $b = 16$, prior strength 64, prior variance .05, gate $\tau = .03$, and margin $\gamma = .05$ across conditions. The no-pilot comparator uses the same supports, optimization rules, engine forecasts, and margin, but keeps the common variance forecast. Comparisons reuse the same trained populations and matching-precision Direct-CS endpoints; selected procedures use independent confirmation streams.

Table 1: **Operational costs and engine selection.** B33/B35: $K_{3,3}/K_{3,5}$; CS: Direct-CS. EB-only/CS reports geometric / arithmetic mean ratios; others are geometric. Ratios below one favor the numerator. Intervals use 30 paired population means: ten-contrast-adjusted for EB-only/Direct-EB, descriptive 95% for CACE-E/CS. All arms complete $1{,}500/1{,}500$ runs. Pilot % is mean pilot charge / mean CS cost. Bold marks routed ratios and the high-error EB excess; tint isolates that contrast. Dashes are unreported. Fixed-allowance/price details: Appendix F.

	EB-only selector					
Setting (ω, t) or η	Opens	/ Direct-EB geom. [adj. interval]	/ CS geom. / mean	No pilot / CS	**CACE-E / CS** geom. [95% interval]	Pilot (%)
A. Primary conditions, $\epsilon = .02$						
$P_4, 0, 1$	8	.6008 [.5937, .6080]	.7553 / .7555	.7559	**.7559** [.7480, .7640]	.00
$P_4, .2, 150$	8	.7087 [.7087, .7088]	.7102 / .7102	.7102	**.7102** [.7101, .7103]	.00
B33, $.1, 0$	64	.3945 [.3909, .3982]	.8742 / .8741	.8742	**.8840** [.8730, .8950]	.47
B35, $.2, 0$	256	.5206 [.5094, .5321]	.9705 / .9705	1.0000	**1.0060** [.9990, 1.0130]	.43
B35, $.4, 0$	256	1.0000 [1.0000, 1.0000]	1.8248 / **1.8256**	1.0000	**1.0050** [.9990, 1.0110]	.43
B. Fixed allowance and opening prices, $\epsilon = .01$						
B33, $\eta = .002$	–	–	1.1401 / –	1.0000	**1.0040** [.9990, 1.0090]	.17
$P_4, \eta = .004$	–	–	1.0052 / –	1.0001	**1.0001** [.9999, 1.0003]	.00

Table 1 reports geometric savings of 24.4%, 29.0%, and 11.6% against Direct-CS in the three useful sharing regimes. Elsewhere, routing predominantly to Direct-CS keeps geometric cost ratios near one, including the tighter-precision cases. The table distinguishes geometric ratios from ratios of arithmetic mean costs; fixed-physical arithmetic comparisons appear in Table 22. All routed conditions complete $1{,}500/1{,}500$ repetitions, with reference brackets of at least $[99.8, 100]\%$.

No-pilot engine choice already avoids the large high-error B35 loss. In homogeneous B33 it is slightly cheaper than paid selection, while the P4 no-pilot branches coincide exactly. Selection information is most useful when the common variance forecast is less representative, as examined below. Figure 3(a,b) summarizes the primary and tighter-precision decisions; paired comparisons with no pilot are reported in Appendix B.7.

Pilot cost and routing margin have separate effects. Always opening the candidate union for a high-setup P4 pilot raises the ratio to 1.001, compared with .7102 under the gate: preparation costs have already been incurred before confirmation. Increasing B33's routing margin from 0 to $.15$ changes the ratio from $.882$ to 1.005, as the rule becomes more conservative. Figure 3(c,d) shows these controls.

6.3 Asymmetric prices, variances, and catalogues

Known price differences affect support and allocation, while pilot observations can improve forecasts of confirmation cost. Four profiles test these effects: configuration-specific observation prices, unequal outcome variances, a non-affine 48-target B33 catalogue with sources outside the list, and unequal P4 opening costs. Three levels use paired populations and fixed profile draws. Let $Z(z)$ be the bounded base score, $V_0(z) = \mathbb{E}Z(z)$, and $\rho_z \in [0, 1]$ control added noise. With a fixed normalized residual profile $|r_z| \le 1$ and $B \mid Z \sim \text{Bernoulli}(Z)$, the variance intervention is

$$Y_\rho(z) = .05 + .9\{(1 - \rho_z)Z(z) + \rho_z B\} + \eta r_z, \tag{19}$$

so $\mathbb{E}Y_\rho = .05 + .9V_0 + \eta r$ while the conditional variance changes. For $Z \in [.05, .95]$ and $\eta \le .01$, outcomes remain bounded in $[0, 1]$. This keeps the assured mean class fixed and exposes only scalar outcomes to the evaluator.

Figure 3(e–h) compares CACE-E with a matched EB-only selector using the same pool, gate, and selection information. At the largest observation-price and opening-price perturbations, routed ratios are .960 and .772. In the latter profile the matched EB-only and routed plans coincide: the improvement comes from price-aware support/allocation decisions. Under high unequal variance

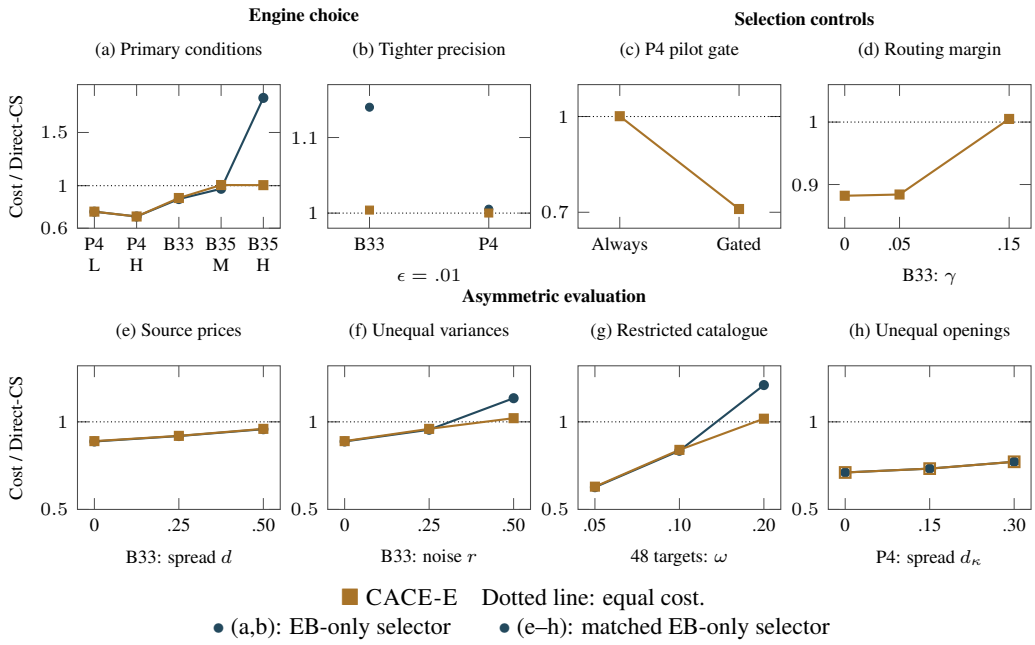

Figure 3: **Engine choice, selection controls, and asymmetric evaluation.** (a,b) Engine choice avoids the largest EB excess; (c,d) pilot opening costs and the routing margin affect savings; (e–h) unequal variances and restricted targets favor engine choice. Ratios include selection cost; below one favors the numerator. The matched EB-only selector shares candidates, gate, and selection information. Bars are descriptive 95% intervals; primary intervals are in Table 1. Open squares mark coincident curves. Top scales differ; bottom scales match. L/H: low/high setup; M/H: moderate/high allowance. Panel (b) fixes allowance/prices; (c) compares union-support pilots with gated selection.

and the high-allowance restricted catalogue, the routed ratios are 1.021 and 1.018, compared with matched EB-only ratios 1.135 and 1.210. Engine choice reduces these excess costs, although both remain above Direct-CS.

Tables 23 and 4 separate support changes from the value of pilot information. On the unequal-variance and restricted-catalogue profiles, paid selection reduces cost by 2.95% and 4.41% relative to the no-pilot selector. In the source-price profile, whose heterogeneity is known before observing outcomes, the no-pilot rule remains slightly cheaper.

6.4 Precision, access, and reference interpretation

Fixed-allowance, fixed-price sensitivities tighten precision without changing the physical problem. The normalized sweeps instead hold (ω, t) fixed and change allowance and prices with ϵ. The supporting EB grid includes 33 conditions cheaper and 37 costlier than Direct-CS; these paired cases describe operating regimes rather than a deployment win rate. Full sweeps and uncertainty are in the appendix.

The observation interface also affects attainable savings. W6 component/replay access reaches precision at 24.576k scalar-equivalent units, compared with 72.609k under scalar-only acquisition. In the external HotpotQA correction study, 2,614 calls reach half-width .1489, while Direct-CS reaches .0972 with 4,465 calls. Only the direct procedure meets the .10 target. Source access and attained precision therefore remain separate comparison criteria; fewer calls alone do not establish a certified saving under matched conditions.

The high-setup P4 coverage bracket uses finer references acquired after the original references left classifications unresolved. Acquisition observations and intervals were unchanged. Appendix E reports both classifications and the additional reference error budget; this refinement concerns evaluation of coverage, not a new confirmation procedure.

7 Related work

Minimum-ℓ_1 prediction and optimal recovery supply reconstruction geometry and the omission obstruction (Bos, 2023; Ettehad & Foucart, 2021; Foucart, 2021). Our contribution is the direct-sample cost boundary and its tensor law under simultaneous precision, distinct from misspecified and transductive bandit objectives (Lattimore et al., 2020; Fiez et al., 2019).

Interaction arrays and G-optimality supply structured designs (Hedayat & Pesotan, 1997; Dey et al., 2005; Kiefer & Wolfowitz, 1960); our exact path envelope additionally prices preparation and includes a nonuniform intermediate support. Mixed-integer design is complementary (Harman & Rosa, 2025).

Coordination training changes policies (Hu et al., 2020; Abderezaei et al., 2025); benchmarks measure team behavior (Zhu et al., 2025). We evaluate frozen policies. Example-efficient evaluation reduces test items (Maia Polo et al., 2024; Wang et al., 2025), while auxiliary predictions and active inference reduce labeling needs (Angelopoulos et al., 2023; Fisch et al., 2024; Zrnic & Candes, 2024). Recent extensions add cross-task calibration (Emmenegger et al., 2026) and anytime-valid inference (Csillag et al., 2025). We combine simultaneous precision with preparation and selection costs, using EB and confidence-sequence engines (Maurer & Pontil, 2009; Howard et al., 2021). Appendix C.1 expands these connections.

8 Discussion

The boundary identifies when sharing can save observations; preparation frontiers identify the cheapest support. Classical design geometry (Kiefer & Wolfowitz, 1960; Dey et al., 2005) therefore needs preparation charges and robust precision: the same model can favor full, compact, or direct sampling as precision and prices change. Tensor composition computes the boundary from catalogue factors. Changing positive observation prices leaves that screen valid but requires new allocations. These optima concern specified estimator and certificate classes; they are not global cost guarantees for every adaptive evaluation procedure.

Engine selection adds a forecast-dependent economic choice. Fresh confirmation protects validity; paid selection must recover its own cost. Known price heterogeneity can be handled before sampling, while uncertain variances or a changed target list can make pilots useful.

For a new catalogue, justify the allowance over its source–target domain, screen supports for estimability and residual width, price feasible plans including assurance and selection, and freeze the confirmation rule. A compact support is economical only when reduced preparation compensates for additional observations.

Combining these evaluation rules with auxiliary-prediction methods requires charging the auxiliary information and preserving confirmation validity under the selected observation law.

Deployment requires measured costs and broader reliability checks (Kapoor et al., 2026; Rabanser et al., 2026). The controlled tasks supply the mean model and allowance; sequential feedback or conjunctive success can violate that span. Obtaining valid assurance may itself erase acquisition savings. For high-setup P4, eight avoided openings offset $4{,}110{,}976 - 91{,}954.76$ extra observations versus Direct-EB only above 502,377.66 scalar-call units per opening. This assigned regime is not measured deployment overhead. Planning runtime and memory scaling are unmeasured; finite-pool search does not resolve exponential catalogue growth. Appendix A.2 details these limits. HotpotQA requires equal attained precision for cost comparisons.

9 Conclusion

The direct boundary, tensor law, and preparation frontiers characterize sharing under a valid model allowance and a fixed certificate. Finite-pool engine selection extends the acquisition comparison while preserving conditional coverage. Practical benefit additionally requires affordable assurance, calibrated preparation costs, and feasible planning computation. The resulting design principle is to select support and inference jointly, and to compare costs at a common precision with every acquired observation charged.

AI USE STATEMENT

Generative AI assisted manuscript proofreading as well as figure and table formatting. The authors are responsible for the final text, claims, and artifacts.

ETHICS STATEMENT

Controlled studies involve no human participants; the external study uses public questions.

REPRODUCIBILITY STATEMENT

Analytic proofs and finite support constructions appear in Appendix C; the acquisition algorithm and confidence procedures are in Appendices A and B. Appendices D, E, F, and G specify the studies and cost accounting. Code for anonymous review is available at `https://anonymous.4open.science/r/compositional-agent-evaluation-8230/`.

REFERENCES

Ava Abderezaei, Chi-Hui Lin, Joseph Miceli, Naren Sivagnanadasan, Stéphane Aroca-Ouellette, Jake Brawer, and Alessandro Roncone. Towards zero-shot coordination between teams of agents: The N-XPlay framework. arXiv:2506.17560, 2025. URL https://arxiv.org/abs/2506.17560.

Abhineet Agarwal, Anish Agarwal, and Suhas Vijaykumar. Synthetic combinations: A causal inference framework for combinatorial interventions. arXiv:2303.14226, 2023. URL https://arxiv.org/abs/2303.14226.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000. URL https://doi.org/10.1126/science.adi6000.

Len Bos. A characterization of optimal prediction measures via ℓ_1 minimization. arXiv:2312.03091, 2023. URL https://arxiv.org/abs/2312.03091.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei A. Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b1041e52d3be19f0a9bc491657488e4a-Abstract-Datasets_and_Benchmarks_Track.html.

Daniel Csillag, Claudio Jose Struchiner, and Guilherme Tegoni Goedert. Prediction-powered E-values. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11493–11514, 2025. URL https://proceedings.mlr.press/v267/csillag25a.html.

Aloke Dey and Chung-yi Suen. Optimal fractional factorial plans for main effects and specified two-factor interactions: A projective geometric approach. *The Annals of Statistics*, 30(5):1512–1523, 2002. doi: 10.1214/aos/1035844986.

Aloke Dey, Chung-yi Suen, and Ashish Das. Asymmetric fractional factorial plans optimal for main effects and specified two-factor interactions. *Statistica Sinica*, 15:751–765, 2005. URL https://www3.stat.sinica.edu.tw/statistica/oldpdf/A15n311.pdf.

Francois K. Domagni, A. S. Hedayat, and Bikas Kumar Sinha. D-optimal saturated designs for main effects and interactions in 2^k-factorial experiments. *Statistical Theory and Related Fields*, 8(3):186–194, 2024. doi: 10.1080/24754269.2024.2341983. URL https://arxiv.org/abs/1911.00794.

Nicolas Emmenegger, Ellery Stahler, and Chara Podimata. Prediction-powered inference across many tasks for AI evaluation & social science research. arXiv:2605.29249, 2026. URL https://arxiv.org/abs/2605.29249.

Mahmood Ettehad and Simon Foucart. Instances of computational optimal recovery: Dealing with observation errors. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1438–1456, 2021. doi: 10.1137/20M1328476. URL https://epubs.siam.org/doi/10.1137/20M1328476.

Tanner Fiez, Lalit Jain, Kevin G. Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8ba6c657b03fc7c8dd4dff8e45defcd2-Abstract.html.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c9fcd02e6445c7dfbad6986abee53d0d-Abstract-Conference.html.

Simon Foucart. Instances of computational optimal recovery: Refined approximability models. *Journal of Complexity*, 62:101503, 2021. doi: 10.1016/j.jco.2020.101503. URL https://www.sciencedirect.com/science/article/pii/S0885064X20300467.

Janet Godolphin. Construction of blocked factorial designs to estimate main effects and selected two-factor interactions. *Journal of the Royal Statistical Society: Series B*, 83(1):5–29, 2021. doi: 10.1111/rssb.12397. URL https://arxiv.org/abs/1907.02373.

Radoslav Harman and Samuel Rosa. Mixed-integer linear programming for computing optimal experimental designs. *Journal of Statistical Planning and Inference*, 234:106200, 2025. doi: 10.1016/j.jspi.2024.106200. URL https://arxiv.org/abs/2305.17562.

A. S. Hedayat and H. Pesotan. Two-level factorial designs for main effects and selected two-factor interactions. *Statistica Sinica*, 2:453–464, 1992. URL https://www3.stat.sinica.edu.tw/statistica/j

2n2/j2n28/j2n28.htm.

A. S. Hedayat and H. Pesotan. Designs for two-level factorial experiments with linear models containing main effects and selected two-factor interactions. *Journal of Statistical Planning and Inference*, 64:109–124, 1997. doi: 10.1016/S0378-3758(96)00209-1.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991. URL https://arxiv.org/abs/1810.08240.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. Other-Play for zero-shot coordination. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4399–4410, 2020. URL https://proceedings.mlr.press/v119/hu20a.html.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=Zy4uFzMviZ.

Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary S. Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndzomga, Dheeraj Oruganty, Sophie Luskin, Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, Juyong Lee, Tengjun Jin, Yifan Mai, Yifei Zhou, Yuxuan Zhu, Rishi Bommasani, Daniel Kang, Dawn Song, Peter Henderson, Yu Su, Percy Liang, and Arvind Narayanan. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.11977.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960. doi: 10.4153/CJM-1960-030-4.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Towards a science of scaling agent systems. arXiv:2512.08296v3, 2026. URL https://arxiv.org/abs/2512.08296v3. Version of April 8, 2026; first preprint December 2025.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5662–5670, 2020. URL https://proceedings.mlr.press/v119/lattimore20a.html.

Daniel LeJeune, Gautam Dasarathy, and Richard Baraniuk. Thresholding graph bandits with GrAPL. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2476–2485, 2020. URL https://proceedings.mlr.press/v108/lejeune20a.html.

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, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. Towards a science of AI agent reliability. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.16666v3. Author version arXiv:2602.16666v3.

Ganghua Wang, Zhaorun Chen, Bo Li, and Haifeng Xu. Cer-Eval: Certifiable and cost-efficient evaluation framework for LLMs. arXiv:2505.03814, 2025. URL https://arxiv.org/abs/2505.03814.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. MultiAgentBench: Evaluating the collaboration and

competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.421. URL https://aclanthology.org/2025.acl-long.421/.

Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62993–63010, 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html.

A ACQUISITION ALGORITHM

Algorithm 1: Engine-aware catalogue evaluation

Inputs: catalogue and source domain; feature matrix; assured allowance η; precision ϵ; budgets δ_A, δ_I; prices (κ_s, c_s); candidate pool; pilot size b; pilot gate τ; routing margin γ; acquisition cap.

1. Screen. Check estimability on each support and require a positive residual-width margin for every omitted target. Apply the analytic direct boundary on its stated domain. Retain feasible direct candidates.
2. Forecast. Construct EB and Direct-CS cost forecasts using their own inference rules, prescribed reporting grids, and common initial variance forecast. Form the candidate union U.
3. Evaluate selection cost. Compute B_U. If $B_U \le \tau \min_p \widehat{C}_p^0$, open U, acquire b observations per source, and update variance forecasts. Otherwise use no pilot and set $U = \varnothing$.
4. Select. Include all sunk selection costs and remaining openings. Choose the cheapest EB plan only if $\widehat{C}_{\mathrm{EB}} \le (1-\gamma)\widehat{C}_{\mathrm{CS}}$; otherwise choose Direct-CS.
5. Freeze. Record the selected engine, support, weights, cumulative schedule, tuning, cap, and simultaneous error allocation. Exclude all selection observations from confirmation.
6. Confirm. Acquire fresh observations under the selected rule until all catalogue intervals have half-width at most ϵ or the cap is reached. An empty confidence intersection causes diagnosis and abstention.
Outputs: simultaneous intervals, completion status, opened sources, consumed counts, and the complete cost ledger.

The algorithm compares a declared finite pool. Its confirmation guarantee is Proposition 8; the exact structural optima retain the assumptions of their respective theorems. Appendix B specifies engine forecasts, caps, and statistical units. Appendix B.7 defines the matched no-selection-pilot control.

A.1 SCOPE OF THE GUARANTEES

Theorem 1, Corollary 2, and the tensor law concern fixed reproducing estimators under the stated Hoeffding certificate and catalogue-only sources; savings below the boundary concern continuous counts. The exact preparation frontiers apply to the four-role path and specified allowance slices. Proposition 8 instead preserves coverage after finite-pool selection, conditional on valid assurance and confirmation. It establishes neither global support optimality nor an oracle-cost guarantee. The controlled scores enforce the interaction span by construction; the external correction study uses a different assurance mechanism.

A.2 SUPPLYING ASSURANCE AND INTERPRETING PRACTICAL COST

A supplied bound is a substantive input. The guarantee requires a bound on mean-model error over the entire declared source–target domain. Small residuals on sampled configurations, a variance prior, or a successful confirmation run do not establish it on unobserved targets. A deterministic structural argument can supply the bound, as in the controlled score construction; otherwise an independent assurance procedure must cover that domain and spend its own error budget.

For example, freeze a fitted mean function $\widetilde{V} = X\widetilde{\beta}$ before an independent audit that returns simultaneous intervals $[L_z, U_z]$ for every domain mean with failure probability at most δ_A. Then

$$\eta_{\mathrm{audit}} = \max_z \max\{|L_z - \widetilde{V}(z)|, |U_z - \widetilde{V}(z)|\}$$

is a valid supplied allowance on the audit event. This is a sufficient workflow, not an assumed procedure in the experiments or a claim that assurance is cheap. Auditing every domain point may approach direct evaluation cost. Charge all assurance observations and any newly required openings before comparing total costs; a large allowance can also remove sharing benefits. Without a valid structural or statistical justification, the sharing guarantee is conditional and direct evaluation remains the model-independent alternative.

What an opening price represents. An opening is a one-time, configuration-specific preparation cost, such as instantiating a workflow or provisioning an isolated environment. The ledger charges each opening once and all consumed observations, including applicable pilots and assurance. Already completed work common to all choices is not a new opening. Assigned prices are in scalar-call units; applying the trade-off requires measured workload-specific preparation and observation costs, rather than an assumed conversion to tokens or latency.

Planning computation and scale. The acquisition ledger excludes candidate construction, conic optimization, and backend runtime; planning-runtime and peak-memory scaling are unmeasured. A binary catalogue has 2^h configurations, and fixed-support planning can require weights for every target–source pair. A finite pool limits search but does not remove this growth or establish a global optimum.

B ENGINE-AWARE EVALUATION: GUARANTEES AND FULL PROTOCOL

B.1 SELECTION AND CONFIRMATION

Let A be the assurance event, with $\mathbb{P}(A^c) \le \delta_A$, and let $\widehat{p}$ be measurable with respect to selection information $\mathcal{F}_0$. Conditional on the selected support, weights, schedule, and tuning, confirmation must use fresh observations from the declared task law. For every realized choice on valid assurance, its failure event $F_{\widehat{p}}$ has conditional probability at most δ_I. Therefore

$$\mathbb{P}(F_{\widehat{p}}) \le \mathbb{P}(A^c) + \mathbb{E}\big[\mathbf{1}_A \mathbb{P}(F_{\widehat{p}} \mid \mathcal{F}_0, A)\big] \le \delta_A + \delta_I.$$

This proves Proposition 8. On a completed catalogue, midpoint-minus-cost decisions have regret at most 2ϵ on the simultaneous coverage event. When assurance itself is random, condition on its recorded output as part of the pre-confirmation information. The argument uses uniform conditional validity on its valid outputs, not an assumption that an unverified fitted residual is correct. Engine multiplicity adds no extra factor because selection precedes confirmation and only one procedure supplies the reported intervals. Within that procedure, all targets and looks still require the declared simultaneous budget. Parallel confirmation followed by an outcome-dependent choice is a different protocol.

For EB, fix support, reproducing weights, cumulative count vectors, and look allocation after selection. A stopped run is the prefix of this fixed schedule; only its stopping time depends on confirmation. For Direct-CS, support and boundary tuning are frozen but the largest-width source choice remains adaptive under the time-uniform guarantee. Neither selection observations nor pilot-based variance estimates are substituted for the confirmation variance bounds. Intersecting intervals over time is permitted only within the valid chosen engine. An empty intersection triggers diagnosis and abstention.

B.2 CANDIDATE POOL, FORECASTS, AND PAID SELECTION

The baseline pool contains matched direct EB, full-support projection, the robust full-support conic candidate, classical product supports, and all distinct feasible solved path supports. The engine-aware pool additionally contains Direct-CS. For asymmetric tasks it also contains the corresponding price-aware allocations and feature-only QR supports and classical-coset unions; no confirmation outcomes are used to generate supports. On a restricted target list, all target rows are reproduced using the declared ambient source domain; a full-cube boundary is not substituted for a domain-specific screen. Candidate aliases share a single specification.

Outcome-free forecasts. Start with $v_0 = .05$ at each source. For a fixed EB candidate, replace sample variance by its forecast in the source-variance upper-bound expression, preserve the cell-variance penalty, and choose the first allowed cumulative checkpoint whose largest forecast radius meets ϵ. Candidate weights are obtained from fixed-support conic planning or the solved construction and then frozen. Count-vector rounding is upward. The existing primary support pool and reporting grids are retained for direct comparisons; asymmetric support/allocation variants are declared before confirmation. The EB per-look budget is $\delta_I/64$ with at most 64 looks, so unused looks do not change a selected bound.

The direct forecast uses the same gamma-exponential boundary function $u(v; \alpha, \rho)$ as confirmation. With its charged, excluded initialization observation and the running-mean predictor, the expected intrinsic variance after n IID observations of variance v_s is $(n + 1 + H_{n-1})v_s$, where H_k is the harmonic sum. Predict the per-source count by the smallest multiple of 64 satisfying

$$u((n + 1 + H_{n-1})\widetilde{v}_s; \alpha, \rho)/n \le \epsilon, \qquad \alpha = \delta_I/(2M).$$

The cap-based tuning rule and the boundary are specified in Appendix B.2. This forecasts the raw boundary, not future intersections or the exact expected stopping cost. The expectation identity is not a concentration bound. Only the confirmation CS determines when the actual procedure finishes.

Pilot gate and forecasting update. A candidate union U includes direct targets. Set $b = 16$, prior strength $\nu = 64$, $\tau = .03$, and $\gamma = .05$. The gate in (17) is evaluated before sampling. If it passes, collect exactly b IID observations per source in U, charge all distinct openings once, and use

$$\widetilde{v}_s = \min\{1/4, \max\{10^{-6}, (b\widehat{v}_s + \nu v_0)/(b + \nu)\}\}.$$

This shrinkage estimate is a forecast only. Recompute engine costs including the sunk selection ledger; choose the cheapest EB candidate, including direct EB, only beyond the fixed relative margin. When the gate fails, use the outcome-free forecast and do not open extra configurations. The Direct-CS engine's own initialization sample remains charged separately. Selection and confirmation random streams are disjoint.

Full ledger and caps. For selected plan p and consumed confirmation counts N_s, the charged cost is

$$C = C_0 + \sum_{s \in S_p \setminus U} \kappa_s + \sum_{s \in S_p} c_s N_s.$$

A setup already paid during selection is not paid again; an unused pilot source is not refunded. Scalar calls, token counts, and wall-clock work are not interchangeable. References and policy training remain separate evaluation-development costs. The displayed selection ledger conditions on supplied assurance with no new assurance acquisitions. If an application must acquire assurance, add that cost to the total comparison and count each opening only once across assurance, selection, and confirmation (Appendix A.2). The total cap is set outcome-free from twice the fixed-Hoeffding direct charge plus the allowable pilot charge. Each inherited comparator endpoint must be checked against that cap before reuse. Noncompletion is charged the cap and never omitted from mean costs.

Direct-CS boundary and tuning. Here we give the implemented rule explicitly. Let $P(a, b)$ denote the regularized lower incomplete gamma function and define

$$\begin{aligned} g(s, v; \rho) = {} & \rho \log \rho - \log \Gamma(\rho) - \log P(\rho, \rho) + \log \Gamma(v + \rho) \\ & + \log P(v + \rho, v + \rho + s) - (v + \rho) \log(v + \rho + s) + s + v, \\ u(v; \alpha, \rho) = {} & \inf\{s \geq 0 : g(s, v; \rho) \geq \log(1/\alpha)\}. \end{aligned}$$

This is the unit-range gamma-exponential mixture boundary of Howard et al. (2021). For a target's observations $Y_1, \ldots, Y_n$, set

$$\widehat{\mu}_{i-1} = \frac{1}{i-1} \sum_{k<i} Y_k \quad (i > 1), \qquad V_n^{\mathrm{pred}} = \sum_{i=1}^{n} (Y_i - \widehat{\mu}_{i-1})^2.$$

The operational procedure initializes $\widehat{\mu}_0$ with one independent, charged observation excluded from the reported mean. Its raw interval is $\overline{Y}_n \pm u(V_n^{\mathrm{pred}}; \alpha, \rho)/n$, with $\alpha = \delta_I/(2M)$ for the two tails and all targets. Intersect these intervals over each target's observation history. Operational intervals are not clipped to $[0, 1]$; an empty intersection causes abstention.

Let N_{cap} be the direct procedure's scalar-observation cap, including its initialization calls and excluding opening charges. The operational and inherited fixed-physical comparisons use

$$\ell_0 = \log(M/\delta_I), \qquad \rho = \frac{N_{\mathrm{cap}}/(8M)}{2\ell_0 + \log(1 + 2\ell_0)}.$$

The cap and ρ are fixed before observations; inherited endpoints keep their original matching-precision tuning when repriced. Sample in batches of 64 from the target with the largest current interval width, breaking ties by target index, and stop at joint precision or the cap. The numerical boundary routine brackets the root by doubling, then uses 42 bisection steps and a 10^{-8} outward offset.

The earlier structured-score study uses the same mixture function with tuning budget $B_0 = \lfloor N_{\mathrm{cap}}/2 \rfloor$ and numerator $B_0/(4M)$ in place of $N_{\mathrm{cap}}/(8M)$. It initializes the predictable mean at $1/2$ and intersects with $[0, 1]$. These protocol differences are retained in their respective comparisons; the external study's block-normalized tuning is specified in Appendix G.3.

B.3 Statistical units, cost estimands, and controls

The five primary routing conditions reuse the original 30 populations per family and 50 repetition labels. The two tighter-precision conditions hold physical allowance and opening price fixed, as in Table 22. Direct-CS endpoints are matching-epsilon inherited acquisitions; new selection and confirmation streams are separate from those endpoints and all earlier acquisitions. Pairing is at the population level, not a claim that independent confirmation streams share every observation.

For population i, average capped cost over its 50 repetitions, then form the log ratio to that population's mean comparator cost. Geometric ratios and t_{29} intervals use these 30 log ratios. All new routing and asymmetry intervals are descriptive 95% intervals; they do not join or replace the corrected original primary family. Write $\overline{C}_{p,i}$ for population i's mean capped cost. The budget comparison and multiplicative paired summary are, respectively,

$$R_A = \frac{\sum_{i=1}^{30} \overline{C}_{p,i}}{\sum_{i=1}^{30} \overline{C}_{\mathrm{CS},i}}, \qquad R_G = \exp\left\{ \frac{1}{30} \sum_{i=1}^{30} \log \frac{\overline{C}_{p,i}}{\overline{C}_{\mathrm{CS},i}} \right\}.$$

The arithmetic ratio estimates relative mean budget under equal population weighting; the geometric ratio describes an average multiplicative comparison and need not equal it. Table 2 reports both for routed conditions. Its intervals belong to R_G only; no arithmetic-ratio interval is inferred from them.

A run is reference-confirmed if all reported intervals contain their independently valid reference intervals, a definite miss if at least one is disjoint, and otherwise indeterminate. Selection must not inspect references, fitted reference residuals, or comparator stopping costs. References supply evaluation diagnostics only. For mean-preserving variance changes, the original reference event can be transformed affinely to the new mean and restricted to the target list. This reuses, rather than multiplies, that event's error budget.

Table 2: Routed costs relative to matching Direct-CS: arithmetic budget ratio R_A, geometric paired ratio R_G, and observed reference classifications. Each row completes $1500/1500$ repetitions with 0 definite misses. Unknown classifications are not counted as demonstrated coverage failures or successes. Their complements form brackets for these recorded runs, not population coverage confidence intervals.

Condition	R_A	R_G	R_G: descriptive 95% interval	Confirmed	Unknown
P_4, nominal	.7561	.7559	$[.7480, .7640]$	1500	0
P_4, high setup	.7102	.7102	$[.7101, .7103]$	1497	3
B33, mild allowance	.8844	.8840	$[.8730, .8950]$	1500	0
B35, moderate allowance	1.0062	1.0060	$[.9990, 1.0130]$	1499	1
B35, high allowance	1.0052	1.0050	$[.9990, 1.0110]$	1498	2
B33, fixed problem, $\epsilon = .01$	1.0042	1.0040	$[.9990, 1.0090]$	1500	0
P_4, fixed problem, $\epsilon = .01$	1.0001	1.0001	$[.9999, 1.0003]$	1499	1

The preparation ablation forces the union-support pilot before reoptimizing the remaining charge; it does not charge a fictitious pilot while keeping compact openings. It yields a high-setup P4 ratio 1.0010 with interval $[.9990, 1.0030]$, compared with the gated ratio .7102. The B33 margin sweep uses $\gamma \in \{0, .05, .15\}$, giving ratios .882, .884, and 1.005, with intervals $[.871, .893]$, $[.873, .895]$, and $[.999, 1.011]$. A conservative selector can forgo a genuine small gain; it is not an ex-post minimum of completed methods.

B.4 Asymmetric profiles and matched information controls

For each population and family, source-specific vectors $u_z \sim \mathrm{Unif}[-1, 1]$ are fixed independently of outcomes and reused across perturbation levels. Profile generators use seed prefix $(20260922, 281828)$ with family and population indices; selection and confirmation append distinct stage tags. Source prices are $c_z = \exp(du_z)/\overline{\exp(du)}$. The B33 price profile has $d \in \{0, .25, .50\}$, $\omega = .1$, and no opening charge. The variance profile uses (19) with $\rho_z = r(1 + u_z)/2$, $r \in \{0, .25, .50\}$, $\omega = .1$, and homogeneous prices. Conditional coin draws are independent of other

acquisition randomness. Although individual noisy outputs need not be pairwise polynomials, their means remain in the supplied approximability class.

The restricted-list profile chooses 48 of B33's 64 targets without replacement using feature-only seed 271828, retains all 64 allowable sources, and varies $\omega \in \{.05, .10, .20\}$. Its target-specific reconstruction screens use that source domain. The P4 opening profile fixes $\omega = .2$, mean opening price $tL = 150L$, and $\kappa_z = \kappa \exp(d_\kappa u_z)/\overline{\exp(d_\kappa u)}$ for $d_\kappa \in \{0, .15, .30\}$. Scalar observation prices remain one. These profiles do not purport to preserve the exact homogeneous support frontiers.

All profiles compare a fixed symmetric EB prescription, a price-aware EB-only pool, CACE-E, and Direct-CS. The EB-only pool has the same source candidates, gate, selection information, cost ledger, and EB confirmation rules as CACE-E, but cannot choose Direct-CS. Exact specification aliases share confirmation evidence. Thus support adaptation and engine choice can be separated. The fixed symmetric prescription is a mechanism control, not the strongest baseline. Parameter levels share base populations and profile draws; they are not additional training replications.

Table 3: **Asymmetric profiles with a matched EB-only control.** Every entry is a geometric ratio to Direct-CS; tint marks the largest level in each profile. Fixed EB retains the symmetric prescription; matched EB adapts support/allocation with the same selection information and charges as CACE-E but cannot change confidence engine. P4 matched-EB and routed plans are evidence aliases. Small differences between matched EB and routing are not claimed as significant advantages. Routing is most consequential when an EB-only choice loses to direct. All descriptive intervals use the same 30-population analysis and exclude no capped failures.

Profile	Level	Fixed EB	Matched EB	CACE-E	Routed 95% interval
Source price, B33	$d = 0$	.880	.888	.890	[.875, .905]
	$d = .25$	.930	.919	.920	[.902, .938]
	$d = .50$	1.050	.958	.960	[.941, .979]
Unequal variance, B33	$r = 0$	.880	.888	.890	[.875, .905]
	$r = .25$	.980	.955	.960	[.942, .978]
	$r = .50$	1.140	1.135	1.021	[1.008, 1.034]
Restricted 48-target list	$\omega = .05$	.610	.627	.630	[.614, .646]
	$\omega = .10$	.810	.835	.840	[.821, .859]
	$\omega = .20$	1.230	1.210	1.018	[1.006, 1.030]
Unequal openings, P4	$d_\kappa = 0$	.710	.711	.711	[.709, .713]
	$d_\kappa = .15$	.748	.733	.733	[.721, .745]
	$d_\kappa = .30$	.821	.772	.772	[.756, .788]

B.5 Operational confidence and conic planning

For $n_s \geq 2$, the source-variance bound used with (16) is

$$\overline{v}_s = \min\left\{\tfrac{1}{4}, \left(\sqrt{\widehat{v}_s} + \sqrt{\frac{2\log(2|S|/\delta)}{n_s - 1}}\right)^2\right\}, \qquad V_x^+ = \sum_s w_{xs}^2 \overline{v}_s / n_s. \tag{20}$$

Use $1/4$ for a source with fewer than two samples. Variance and mean-error events split the failure budget as proved in Appendix B.6. The point estimate used for later decisions is the midpoint of the final valid interval; range intersection and midpoint selection must not silently be confused with the original untruncated linear estimate.

For a declared support, the continuous fixed-width optimization has the exact reformulation

$$\begin{aligned} \min_{n,u,\tau} \quad & \sum_{s\in S} c_s n_s \\ \text{subject to} \quad & X_S^\top u_x = \tau_x x_x, \qquad \tau_x \geq 0, \quad n_s \geq 0, \\ & \eta \left\|\tau_x e_x - J u_x\right\|_1 + 1 \leq \epsilon \tau_x, \qquad A \sum_{s\in S} u_{xs}^2 / n_s \leq 1. \end{aligned} \tag{21}$$

The recovered weights are $w_x = u_x/\tau_x$. The affine-norm inequality ensures $\tau_x > 0$; the quadratic-over-linear terms are closed perspectives, with $0^2/0 = 0$ and nonzero-over-zero infeasible. The change of variables and upward-rounding proof are in Appendix B.6. This optimizes the Hoeffding certificate for a fixed support; it neither convexifies the opening charge nor establishes empirical-Bernstein global optimality.

B.6 Confidence, optimization, and reuse details

For a reproducing row, $V_x - w_x^\top V_S = (e_x - Jw_x)^\top r$, bounded by $\eta \|e_x - Jw_x\|_1$. Independent bounded source-mean errors have squared-range sum $\sum_s w_{xs}^2/n_s$. Weighted Hoeffding and a union bound over targets prove (3), including negative weights and source–target overlap.

For (20), give each source variance bound failure probability $\delta/(2|S|)$. Standard-deviation concentration (Maurer & Pontil, 2009) and the deterministic variance bound $1/4$ yield a joint upper variance event. Give the remaining $\delta/2$ to two-sided Bernstein bounds over M targets. The largest individual weighted increment is $b_x = \max_s |w_{xs}|/n_s$; solving the Bernstein quadratic yields $a_x + \sqrt{2V_x^+\ell + a_x^2}$ with $a_x = b_x\ell/3$. No independence between the estimated variance event and the mean error is required: event containment and a union bound suffice. Counts and weights must still be fixed independently of their confirmation outcomes.

For a finite desired-width plan set $g_x = \epsilon - \eta \|e_x - Jw_x\|_1 > 0$, $\tau_x = 1/g_x$, and $u_x = \tau_x w_x$. This gives (21). Conversely its affine constraint forces $\tau_x > 0$; divide by τ_x to recover bias plus noise at most ϵ. The convention $0^2/0 = 0$ and nonzero-over-zero equal to infinity gives the closed perspective. Upward integer rounding decreases noise and adds less than one observation per active source.

A deterministic schedule of J cumulative looks can use δ/J per look without independence across looks. If support or weights depend on prior outcomes, a fresh confirmation stage j can use $\delta_j = \delta/[(j+1)(j+2)]$; conditional validity integrates to joint coverage because the budgets sum to δ. Alternatively, simultaneous cell confidence sequences can be intersected with the supplied model class and projected to target values under adaptive acquisition. An empty intersection requires diagnosis rather than a fabricated narrow interval.

On $\left\|\widehat{V} - V\right\|_\infty \le \epsilon$, let x^* maximize $V - d$ over any feasible subset and $\widehat{x}$ maximize $\widehat{V} - d$. Then

$$V(x^*) - d_{x^*} - V(\widehat{x}) + d_{\widehat{x}} \le 2\epsilon + [\widehat{V}(x^*) - d_{x^*}] - [\widehat{V}(\widehat{x}) - d_{\widehat{x}}] \le 2\epsilon.$$

This is a pathwise reuse result for the same catalogue, not an independent guarantee for new tasks. If a team-independent pre-outcome exception gate is available, $V = (1-p)V_0 + pV_1$ with structured V_0 and bounded V_1 admits pooled allowance $p/2$ when the model contains an intercept. Unknown prevalence and missed exceptions require their own error budgets; fitting a gate from target residuals does not supply assurance.

B.7 Matched no-pilot comparison

The no-pilot selector uses the same candidates, optimizers, engines, caps, routing margin $\gamma = .05$, and initial variance $v_0 = .05$ as CACE-E, but disables its optional selection pilot. Direct-CS retains its charged initialization observation. Thus the comparison isolates selection information while preserving known prices, source access, and engine forecasts.

The seven primary/fixed-physical conditions and the largest level of each asymmetric profile reuse the same 30 populations. Identical frozen specifications share confirmation evidence, including the gated P4 no-pilot branches; distinct specifications use fresh confirmation under the same reference event. Inherited endpoints must match the method-specific cap and Direct-CS tuning.

For each population, average capped cost across its 50 repetitions and then take the log ratio. If G_R and G_N are the geometric routed and no-pilot ratios to the same Direct-CS population means, their paired ratio is exactly G_R/G_N. The uncertainty interval is computed from the paired log differences, not from division of marginal interval endpoints. All comparisons in this appendix are descriptive 95% intervals and remain separate from the original corrected primary contrast family.

The matched comparison separates engine availability from pilot information. The no-pilot selector already removes the high-error engine-blind loss. In the homogeneous B33 condition, paid selection

Table 4: Full selection-information comparison. Ratios include all method-specific selection and preparation costs. Lower paired ratios favor acquiring selection information; tint and bold identify the variance/catalogue contrasts. Exact P4 ties are shared-specification aliases rather than independent measurements with coincident means. The four asymmetric rows use their largest tested perturbation/allowance level.

	Geometric cost / Direct-CS		CACE-E / no pilot	
Condition	No pilot	CACE-E	Paired ratio	95% interval
P_4, nominal	.7559	.7559	1.000000	[1.000, 1.000]
P_4, high setup	.7102	.7102	1.000000	[1.000, 1.000]
B33, mild allowance	.8742	.8840	1.011210	[.999, 1.024]
B35, moderate allowance	1.0000	1.0060	1.006000	[.999, 1.013]
B35, high allowance	1.0000	1.0050	1.005000	[.999, 1.011]
B33, tighter precision	1.0000	1.0040	1.004000	[.999, 1.009]
P_4, tighter precision	1.0001	1.0001	1.000000	[1.000, 1.000]
B33, source prices	.953	.960	1.007345	[.995, 1.020]
B33, unequal variances	1.052	1.021	**.970532**	[.955, .987]
48-target catalogue	1.065	1.018	**.955869**	[.941, .971]
P_4, unequal openings	.772	.772	1.000000	[1.000, 1.000]

has a small excess cost whose paired interval includes one. Under unequal variances and the restricted catalogue, ratios .970532 and .955869 locate a more useful role for selection information. The observation-price profile has known heterogeneity and a paid/no-pilot ratio 1.007345; the preparation-heavy opening-price profile uses the same no-pilot branch under both rules.

An interval containing one is not an equivalence test.

C Geometry and proofs

C.1 Connections to prior work

Selected-interaction factorial designs, projective constructions, and product arrays supply classical supports (Hedayat & Pesotan, 1992; 1997; Dey & Suen, 2002; Dey et al., 2005; Godolphin, 2021; Domagni et al., 2024); optimal-design theory relates information matrices to prediction precision (Kiefer & Wolfowitz, 1960; Pukelsheim, 2006). Mixed-integer design provides complementary optimization tools (Harman & Rosa, 2025). Minimum-ℓ_1 prediction and optimal recovery supply the reconstruction duality used below (Bos, 2023; Ettehad & Foucart, 2021; Foucart, 2021); bandit formulations address different terminal decisions (Lattimore et al., 2020; Fiez et al., 2019; LeJeune et al., 2020). Coordination training changes policies (Hu et al., 2020; Abderezaei et al., 2025), and Synthetic Combinations studies combinatorial interventions (Agarwal et al., 2023); our target is evaluation of frozen policies. The agent-reliability setting is broader (Rabanser et al., 2026). Example-efficient evaluation and auxiliary-prediction methods change the acquired information (Maia Polo et al., 2024; Angelopoulos et al., 2023; Fisch et al., 2024; Zrnic & Candes, 2024). EB bounds, confidence sequences, and convex perspectives are inherited components (Maurer & Pontil, 2009; Howard et al., 2021; Boyd & Vandenberghe, 2004).

C.2 Catalogue restriction and optimal precision

An affine catalogue has the duplicate-free parameterization $z_i(u) = (-1)^{c_i + \ell_i^\top u}$, $u \in \mathbb{F}_2^r$, with $\operatorname{rank}(L) = r$. For original interaction masks $\mathcal{A}_G = \{0, e_i\} \cup \{e_i + e_j : ij \in E_G\}$, restriction gives

$$\mathcal{B} = \{L^\top a : a \in \mathcal{A}_G\}, \quad d = |\mathcal{B}|, \quad \psi(u) = (\chi_b(u))_{b \in \mathcal{B}}, \quad \chi_b(u) = (-1)^{b^\top u}. \tag{22}$$

Coincident masks combine with coefficients $\beta_b = \sum_{a : L^\top a = b} (-1)^{a^\top c} \theta_a$. Character orthogonality gives $M^{-1} \sum_u \psi(u)\psi(u)^\top = I_d$ and $\|\psi(u)\|_2^2 = d$. A face with free roles D has $d = 1 + |D| + |E_G(D)|$. Non-affine lists use the feature matrix and rank directly. The symbol L in this parameterization is the catalogue matrix only; in cost formulas it denotes A/ϵ^2.

For the information matrix and prediction factor in (11), the worst linear-unbiased prediction variance under homogeneous noise is $\sigma^2\Gamma/N$.

Proposition 9 (Feature products determine optimal precision)**.** *Let $\Delta = \{b + b' : b, b' \in \mathcal{B}, b \neq b'\}$. The minimum factor is d, attained exactly when*

$$I(\pi) = I_d \quad \Longleftrightarrow \quad \sum_u \pi_u \chi_a(u) = 0 \quad \textit{for every } a \in \Delta. \tag{23}$$

At least d distinct configurations are needed for finite-variance estimation unbiased over the model.

The trace argument gives $\operatorname{tr} I^{-1} \geq d$ because $\operatorname{tr} I = d$; equality requires $I = I_d$. Off-diagonal entries are feature-product moments, and rank gives the support floor. Thus optimal precision can require balancing interactions absent from the mean model itself.

Theorem 10 (Same dimension, different optimal support)**.** *The four-role star $K_{1,3}$ and path P_4 have $d = 8$, $M = 16$, and optimum factor eight, but their minimum supports attaining it contain eight and sixteen configurations, respectively.*

The star's leaf-parity half-cube balances the required moments. The path's feature products contain every nonconstant four-bit character, forcing the uniform full cube. This is a precision distinction, not an identifiability distinction. For degree-at-most-k features on a full h-bit cube with $h > 2k$, the even-parity half-cube is already information-optimal; Appendix C.6 proves the scope result.

C.3 Scalable classical reference

For $K_{a,b}$, the classical product design uses $s = q_A q_B$, where $q_A = 2^{\lceil \log_2(a+1) \rceil}$ and $q_B = 2^{\lceil \log_2(b+1) \rceil}$, with $d = (a+1)(b+1)$ and $I = I_d$ (Dey et al., 2005). Therefore $d \leq s < 4d$. Equal counts $n = \lceil Ld/s \rceil$ cost $\kappa s + cs\lceil Ld/s \rceil$; the lower bound $\kappa d + cLd$ gives an approximation factor at most four for $L \geq 2$. This construction is shared with the classical comparator. Its cost bound and Corollary 5 answer different questions: affordable exact-model preparation versus the allowance beyond which sharing ceases to help.

C.4 Exact bounded-mean ambiguity at an omitted coordinate

We prove Proposition 3 using finite-dimensional reconstruction duality and optimal recovery (Bos, 2023; Ettehad & Foucart, 2021; Foucart, 2021).

Let $\mathcal{Q} \subseteq \mathbb{R}^M$ be a linear subspace containing the constant vector and let x be a fixed excluded target. Define

$$\mathcal{V}_\eta = \{V \in [0,1]^M : V = q + r,\ q \in \mathcal{Q},\ \|r\|_\infty \leq \eta\}.$$

When omitted-target reconstruction is possible, finite-dimensional linear-program duality gives an attained optimizer

$$g \in \mathcal{Q}, \qquad g(x) = \Lambda_x, \qquad |g(s)| \leq 1 \quad (s \neq x).$$

Because the model contains constants, every reproducing vector sums to one and $\Lambda_x \geq 1$. Put $\alpha = \min\{\eta, [2(1+\Lambda_x)]^{-1}\}$ and define

$$q^\pm = \tfrac{1}{2}\mathbf{1} \pm \alpha g, \qquad r_s^\pm = \mp \alpha g(s)\ (s \neq x), \qquad r_x^\pm = \pm\alpha.$$

Both $q^\pm$ belong to $\mathcal{Q}$ and $\|r\|_\infty^\pm \leq \eta$. Their means satisfy

$$V_s^\pm = \tfrac{1}{2} \quad (s \neq x), \qquad V_x^\pm = \tfrac{1}{2} \pm \alpha(1+\Lambda_x) = \tfrac{1}{2} \pm D_x(\eta).$$

All coordinates are in $[0,1]$, including at the clipping boundary. This constructs both extremes under the full bounded-mean class, rather than using unbounded counterfactual rewards.

Conversely, let v be a minimum-norm omitted-target representation. Intercept reproduction gives $\sum_s v_s = 1$. For any $V \in \mathcal{V}_\eta$ with $V_s = 1/2$ off target,

$$\left|V_x - \tfrac{1}{2}\right| = |V_x - v^\top V_{-x}| = |r_x - v^\top r_{-x}| \leq \eta(1+\Lambda_x).$$

The deterministic range bound also gives a radius at most $1/2$. Thus the displayed extremes are sharp. Convexity of $\mathcal{V}_\eta$ and linearity of the observed-coordinate restriction imply that every intermediate

target value is feasible. In particular, D_x is the exact deterministic minimax recovery error at the observed vector $\frac{1}{2}\mathbf{1}$, not a conjectured confidence radius.

If $\Lambda_x = \infty$, the target evaluation functional is not in the span of the other evaluation functionals. Linear algebra therefore supplies $g \in \mathcal{Q}$ with $g(x) = 1$ and $g(s) = 0$ off target. The two exact-model means $V^{\pm} = \frac{1}{2}\mathbf{1} \pm \frac{1}{2}g$ agree on every available source and have target values zero and one. This remains true at $\eta = 0$ and avoids interpreting $0 \cdot \infty$ in (8).

Sequential confidence and completion. Choose independent bounded Bernoulli outcome laws with means V^+ and V^-. At every allowed source $s \neq x$ the laws coincide. Even an adaptive acquisition policy, randomized estimator, or sequential stopping rule therefore has the same data-transcript distribution in the two worlds, provided it never evaluates the fixed excluded target. Couple the internal randomization as well.

Let C be the event that a complete catalogue with half-width at most ϵ is issued at some admissible look, and choose the first such look. If $\epsilon < D_x$, the two target endpoints are more than 2ϵ apart. At that look the common interval cannot contain both endpoints. Let F^+ and F^- be failure of target coverage at any admissible look in the two respective worlds. Under the common transcript law,

$$C \subseteq F^+ \cup F^-, \qquad \mathbb{P}(C) \leq \mathbb{P}(F^+) + \mathbb{P}(F^-) \leq 2\delta.$$

Uniform simultaneous coverage in (2) implies the last bounds for each admissible mean. The claim also applies with a finite reporting horizon. It does not prohibit abstention: it quantifies why abstention cannot be counted as successful certification. It does not establish a universal lower bound on the number of direct samples once x is permitted, and it does not automatically address a random omitted target selected differently across transcripts.

For $\epsilon < 1/2$ and finite Λ_x, the obstruction holds whenever $\eta/\epsilon > 1/(1 + \Lambda_x)$. At equality this construction has separation 2ϵ rather than strictly greater; no arbitrary-inference impossibility at equality is claimed. The fixed finite-noise certificate theorem includes equality for a separate reason. At $\omega > \omega_*$ every fixed target omission is obstructed in this worst-case sense, while the per-target requirement $n_x \geq L$ remains restricted to (4).

C.5 Support-specific screening before count optimization

Fix a support S such that every target feature row is in the span of X_S. For $x \notin S$, let $\Lambda_x(S) = \min\{\|v\|_1 : X_S^\top v = x_x\}$. For any reproducing estimator on S, the residual term is at least $\eta(1 + \Lambda_x(S))$. The intercept forces a nonzero weight vector; finite positive counts then give a strictly positive Hoeffding noise term. Consequently finite-count feasibility requires (7) at every omitted target.

For sufficiency, choose a minimum-norm representation for each omitted target and a direct row for each included target. Every row has a positive remaining noise margin. There are finitely many rows and each has a finite squared weight norm. A sufficiently large common count at every source drives all noise terms below their respective margins simultaneously. Thus the strict screening rule is necessary and sufficient for existence of a finite-count plan within this certificate class. If estimability fails, reject the support before this step, including when $\eta = 0$. A full support has no omitted targets and the screen is vacuous; direct sampling supplies a feasible plan.

The screen does not give a total-cost optimum or prove that the surviving support improves on direct. It is a cheap logical prerequisite to comparing those costs. The distinction matters at P_4, $\omega = 1/5$: the nominal eight-source plan is rejected because an omitted residual multiplier is five, whereas a different eight-source support with multiplier four remains feasible. Candidates failing this outcome-free screen are reported as infeasible and have no acquisition endpoints.

C.6 Catalogue restriction and the parity scope result

For an original interaction mask a, substitution into the affine catalogue gives $\chi_a(z(u)) = (-1)^{a^\top c}\chi_{L^\top a}(u)$. Grouping equal masks yields (22). Every distinct restricted frequency has an original representative, so its coefficient can vary independently and the dimension is exactly d. Orthogonality of binary characters proves the uniform Gram identity. On a face, each interaction reduces to a constant, a free-role main effect, or an edge between two free roles.

For completeness, the trace bound in Proposition 9 follows from positive eigenvalues λ_j of I: $\sum_j \lambda_j = d$ and $\sum_j \lambda_j^{-1} \geq d$, with equality only when each eigenvalue is one. A rank-deficient information matrix leaves an unobserved coefficient direction that changes some catalogue target. For equal-price, homogeneous-noise ambient sources, select one original monomial representative per target frequency. Every source row of that submodel still has squared norm d, while target rows are orthogonal, so the same lower bound applies within the linear-unbiased comparison. It does not apply to arbitrary biased estimators or informative parameter constraints.

Proposition 11 (A degree barrier to full-support necessity)**.** *On a full h-bit cube, suppose every model feature is a character of degree at most k and $h > 2k$. Uniform evaluation on the even-parity half-cube gives $I = I_d$.*

Proof. Let v be the all-ones frequency mask. The even-parity measure has probability $2^{-h}(1+\chi_v(z))$ at each cube point. Its only nonzero Fourier moments are at 0 and v. The product of two distinct degree-at-most-k characters has a nonzero mask of size at most $2k < h$, and hence cannot equal v. All off-diagonal information entries vanish and diagonal entries equal one. □

For a five-role path this gives a 16-source optimum with rank ten and 32 targets. For the complete pairwise model on five roles, rank is sixteen and the same half-cube is saturated. The proposition supplies a feasible optimal-variance support, not a minimum-support theorem for every graph or restricted affine catalogue.

C.7 Proof of the direct boundary

For a reproducing row put $a = w_{xx}$. If $\Lambda_x < \infty$, then $X_{-x}^\top w_{x,-x} = (1-a)x_x$, and homogeneity of the minimum-ℓ_1 problem gives

$$\|e_x - w_x\|_1 = |1-a| + \|w_{x,-x}\|_1 \geq (1+\Lambda_x)|1-a|.$$

If $\omega(1+\Lambda_x) \geq 1$ and $q_x < 1$, every $a \neq 0$ has radius at least $|1-a| + |a|/\sqrt{q_x} > 1$. If $a = 0$, the bias term is at least one; reproduction of the intercept forces a nonzero off-target weight and positive finite-count noise. Thus this case is also infeasible, including omission of the target. If $\Lambda_x = \infty$, any $a \neq 1$ would give an omitted-target representation after division by $1-a$, so $a = 1$ necessarily and $q_x \geq 1$ follows. These arguments establish the direct lower bound for every target when $\omega \geq \omega_*$; direct rows attain it. Integer counts then require and attain $\lceil L \rceil$ per target.

For the converse, choose a target with $\omega(1+\Lambda_x) < 1$ and an optimal omitted-target vector v. Keep all other rows direct with counts $q_s = 1$. Set $w_{xx} = 1-\alpha$, $w_{x,-x} = \alpha v$, and $q_x = 1 - b\alpha$, where $0 < b < 2[1-\omega(1+\Lambda_x)]$. The normalized radius becomes

$$\omega\alpha(1+\Lambda_x) + \sqrt{\frac{(1-\alpha)^2}{1-b\alpha} + \alpha^2 \|v\|_2^2}.$$

At $\alpha = 0$ it equals one and its derivative is $\omega(1+\Lambda_x) - 1 + b/2 < 0$. A sufficiently small positive α therefore preserves feasibility and lowers the observation count without changing opening costs. This proves the continuous iff statement. Rounding can erase a small continuous saving, so an integer converse is not asserted.

Heterogeneous prices. Corollary 2 follows from the per-coordinate lower bound $n_x \geq L$ in Theorem 1. Summing with $c_x > 0$ and adding nonnegative setup charges proves the direct lower bound for arbitrary known prices. Below the boundary, the same perturbation of a single direct row decreases only n_x, leaves all other counts and openings fixed, and strictly decreases cost by its positive price. The integer optimum above the boundary is $\sum_x \kappa_x + \lceil L \rceil \sum_x c_x$. This proof assumes the same radius, range, precision, and confidence allocation; heterogeneous noise engines do not inherit the count formula.

C.8 Tensor composition: full proof and analytic bipartite family

The omitted-target linear program has the dual representation

$$\Lambda_x = \sup\{g(x) : g \in \mathcal{Q},\ |g(s)| \leq 1 \text{ for all } s \neq x\}. \tag{24}$$

This is finite-dimensional linear-program duality for reproducing coefficients, consistent with the optimal-prediction geometry of Bos (2023). If finite, the optimum is attained and is at least one because the model contains constants. If omitted-target reproduction is impossible, separation supplies a function $d_x \in \mathcal{Q}$ with $d_x(x) = 1$ and $d_x(s) = 0$ at every other catalogue point.

Consider two factors. For any finite $\Lambda^{(1)}_{x_1}$, use its optimal weights on points (s_1, x_2) with $s_1 \neq x_1$. For every product $f_1 f_2$, those weights reproduce $f_1(x_1)f_2(x_2)$, so they reproduce the whole tensor span and have the same absolute sum. The second factor gives the analogous construction. Therefore $\Lambda^{\otimes}_{(x_1,x_2)} \leq \min(\Lambda^{(1)}_{x_1}, \Lambda^{(2)}_{x_2})$ whenever at least one value is finite.

When both values are finite, denote them by λ_1, λ_2 and choose dual optimizers g_1, g_2. The function

$$g(s_1, s_2) = \frac{g_1(s_1)g_2(s_2)}{\max(\lambda_1, \lambda_2)}$$

belongs to the tensor span. At a non-target point at least one coordinate differs from its target. Its corresponding factor has absolute value at most one; the other has absolute value at most its target optimum, because $\lambda_j \geq 1$. Thus $|g| \leq 1$ away from the joint target, while $g(x_1, x_2) = \lambda_1\lambda_2/\max(\lambda_1, \lambda_2) = \min(\lambda_1, \lambda_2)$. Duality proves the reverse inequality. If only the first factor is nonreproducible, use $d_{x_1} g_2$ to attain λ_2; the symmetric case is identical. If both are nonreproducible, $d_{x_1} d_{x_2}$ isolates the joint target, which is likewise nonreproducible. Induction proves the norm identity for any finite number of factors. Finally,

$$\max_{x_1,\ldots,x_J} \frac{1}{1 + \min_j \Lambda^{(j)}_{x_j}} = \max_j \max_{x_j} \frac{1}{1 + \Lambda^{(j)}_{x_j}},$$

proving the threshold identity.

Additive binary factors. By sign symmetry choose the all-minus target on $m \geq 2$ roles. Assign weight $1/(m-1)$ to each of the m single-flip points and weight $-1/(m-1)$ to the all-flip point. The weights sum to one. Each coordinate has weighted sum $[(2-m)-1]/(m-1) = -1$, so the target's intercept and main effects are reproduced. The absolute sum is $(m+1)/(m-1)$. If j coordinates are flipped, the affine function $g = (m+1-2j)/(m-1)$ takes that value at the target and remains in $[-1, 1]$ for $1 \leq j \leq m$. It is a matching dual witness. For $m = 1$, intercept and main effect span both values, so omission destroys estimability.

The $K_{a,b}$ feature space factors as $\operatorname{span}\{1, z_1, \ldots, z_a\} \otimes \operatorname{span}\{1, z_{a+1}, \ldots, z_{a+b}\}$. The norm formula follows from the tensor theorem and decreases with block size. This proves Corollary 5, including one-role blocks and the saturated $K_{1,1}$ case. The star threshold is recovered as $K_{1,m}$, rather than extrapolated from numerical examples.

C.9 Explicit path designs and nominal support exclusion

Configuration IDs list $(-1, +1)^4$ lexicographically, with the first role most significant and ID zero equal to all minus. The nominal eight-point support is

$$S_8 = \{0, 3, 5, 7, 9, 10, 12, 14\}.$$

Uniform mass gives prediction factor sixteen. Writing roles as (a, b, c, d), the path can be reparameterized as $q(a, b, c, d) = \mu_{bc} + a u_b + d v_c$. The support retains $ad = H_{bc}$ for $H = \left(\begin{smallmatrix} 1 & -1 \\ -1 & -1 \end{smallmatrix}\right)$, with $HH^\top = 2I_2$. Its target factors are $12 - 4adH_{bc}$: eight at sampled points and sixteen at omitted points. At target ID 1, the unique reproducing coefficients have eight absolute values $1/2$. For any probability allocation on this support, Cauchy–Schwarz therefore gives variance factor at least $(8/2)^2 = 16$.

The nominal twelve-point support is

$$S_{12} = \{0, 2, 3, 4, 5, 7, 9, 10, 11, 12, 13, 14\}.$$

IDs $0, 7, 9, 14$ receive mass $\alpha = (\sqrt{2} - 1)/4$; the other eight receive $\beta = (2 - \sqrt{2})/8$. The omitted-target factor is

$$g(\alpha) = \frac{1}{2\alpha} + \frac{4}{1 - 4\alpha} \geq 6 + 4\sqrt{2}.$$

For the symmetric allocation equality holds at the displayed masses. To prove the bound for arbitrary allocations, write $G_* = 6 + 4\sqrt{2}$, $z_j = I_*^{-1} x_j$ for omitted targets $j = 1, 15$, and assign target weights $1/2$ each. Symbolic multiplication yields $x_j^\top z_j = G_*$ and $[(x_s^\top z_1)^2 + (x_s^\top z_{15})^2]/2 = G_*$ for every source $s \in S_{12}$. For arbitrary nonnegative target weights λ_j summing to one, matrix Cauchy–Schwarz gives the valid design bound

$$\Gamma(\pi) \geq \frac{(\sum_j \lambda_j x_j^\top z_j)^2}{\max_{s \in S} \sum_j \lambda_j (x_s^\top z_j)^2}. \tag{25}$$

The displayed witness proves the arbitrary-allocation optimum. There are eight twelve-point supports of this structural form, in two symmetry classes.

Exact global exclusion. Sign changes and path reversal preserve the feature Gram matrix and generate 2,302 disjoint support orbits covering all 65,536 subsets. Of the 1,379 orbits with at least eight points, exact ranks exclude 516. For each of the 861 ordinary full-rank orbits, rational witnesses in (25) establish $\Gamma + |S| \geq 24$. The two remaining orbits are the exceptional twelve-point cases. For ordinary supports, $\Gamma + t|S| \geq 24 + (t-1)|S|$, which is at least $8 + 16t$ for $t \leq 1$ and $16 + 8t$ for $t \geq 1$. Combining the exceptional bound and the three attaining designs proves (13) for all $t \geq 0$. Numerical optimizers propose witnesses; rational arithmetic, exact ranks, and orbit coverage certify this conclusion.

C.10 Robust sampling curve and positive-setup slice

For the path, the polynomial

$$g(z) = \tfrac{1}{2}(1 - z_1 - z_4 + z_1 z_2 + z_2 z_3 + z_3 z_4)$$

equals three at the all-minus target and has absolute value at most one elsewhere. A three-term signed omitted-target representation attains norm three; sign changes transport both witnesses. This proves $\Lambda_x = 3$ for every target.

Let $P = XX^\top/16$. Walsh orthogonality gives $P^2 = P$, $P_{xx} = 1/2$, and $\|e_x - P_{x,:}\|_1 = 2$. For $W = (1-\gamma)I + \gamma P$, the residual norm is 2γ and squared row norm is $1 - \gamma + \gamma^2/2$. At $\gamma = (1-4\omega)/(1-2\omega)$,

$$\frac{1 - \gamma + \gamma^2/2}{(1 - 2\omega\gamma)^2} = \frac{1}{1 + (1-4\omega)^2},$$

which proves feasibility of the claimed equal-count sampling curve.

For its lower bound, use the homogeneous convex constraints

$$u_j^\top X = \tau_j x_j^\top, \qquad \omega \|\tau_j e_j - u_j\|_1 + 1 \leq \tau_j, \qquad \sum_s u_{js}^2/q_s \leq 1. \tag{26}$$

These are equivalent to (4) by $u_j = \tau_j w_j$. Averaging over coordinate sign changes preserves the constraints and sample cost and yields equal $q_s = q$; it need not preserve a sparse support's opening cost. For target zero, any $\|v\|_\infty \leq \omega$ and $x_0^\top b + v_0 = 1$ give $1 \leq (v + Xb)^\top u \leq \|v + Xb\|_2 \|u\|_2$. Write the displayed path polynomial as $g = Xy$ and choose $v = \omega(4e_0 - g)$ and $b = (1-4\omega)x_0/8 + \omega y$. Then

$$v + Xb = e_0 + (1 - 4\omega)(2Pe_0 - e_0).$$

The two vectors on the right are orthogonal unit vectors, so $q \geq \|u\|_2^2 \geq [1 + (1-4\omega)^2]^{-1}$. This gives the matching total-count lower bound for $0 \leq \omega \leq 1/4$; the direct theorem handles larger allowances.

Attaining supports at $\omega = 1/5$. The full plan has $W = 2I/3 + P/3$ and $q_s = 25/26$, total $200/13$. On S_{12}, use nominal reproducing weights and counts $25(1+\sqrt{2})/2$ on the four heavy points and $25(2+\sqrt{2})/4$ on the remaining eight. The total is $150 + 100\sqrt{2}$; residual norms are at most four and omitted active targets have noise margin $1/5$. The nominal dual restricted to those omitted targets, whose minimum norm is three, gives the matching factor-25 lower bound.

The robust eight-point support is

$$S_8^R = \{0, 1, 2, 4, 6, 7, 10, 12\}, \qquad W = XX_{S_8^R}^{-1}, \qquad q_s = 75.$$

Every omitted row contains exactly three coefficients in $\{-1, +1\}$, giving residual norm four and noise $\sqrt{3/75} = 1/5$. Each included target has its direct row. Every source occurs in three omitted rows; summing their constraints gives $3\sum_s q_s^{-1} \le 8/25$. Since $(\sum_s q_s)(\sum_s q_s^{-1}) \ge 64$, total count is at least 600, attained by the displayed equal counts. The nominal S_8 instead has omitted-target absolute sum four and residual norm five, making it infeasible at any finite count when $\omega = 1/5$.

Exact robust exclusion. The same 2,302 orbits contain 516 rank-deficient eligible classes and 514 further classes with a dual omitted-target norm at least four, which exhausts the error budget. For the remaining 349 classes, norm witnesses give target-specific noise margins. Scaled nominal duals provide sample lower bounds; unresolved classes use robust dual witnesses. If $\|v_j\|_\infty \le 1/5$, $x_j^\top b_j + v_{jj} = 1$, and $A_{js} = v_{js} + x_s^\top b_j$, then (26) implies $1 \le \sum_s A_{js} u_{js}$ and hence $\sum_s A_{js}^2 q_s \ge 1$. For nonnegative multipliers α_j this yields

$$\sum_s q_s \ge \frac{\sum_j \alpha_j}{\max_{s \in S} \sum_j \alpha_j A_{js}^2}. \tag{27}$$

Fifteen retained and 45 additional rational dual certificates complete the exclusions. Six attaining symmetry classes contain one full support, eight twelve-point supports, and 32 robust eight-point supports. Each certified lower-bound line is compared with the asserted envelope at $0, t_A, t_B$ in $\mathbb{Q}(\sqrt{2})$. Linearity on the bounded phases and slope at least eight after t_B prove the complete ray. This establishes the global continuous frontier, not global integer optimality or other allowance slices.

C.11 Analytic costs and finite-sample stress

Table 5: **Analytic preparation costs.** All denominators use the same fixed Hoeffding certificate; tint/bold mark twelve-source optima. In the first row, the best eight- or sixteen-source shared design costs 24, rather than the direct cost 32. Setup prices are assigned, not measured deployment overheads.

Analytic certificate regime	t	Opens	$C^*/(cL)$	Direct cost
Exact path, $\omega = 0$	1	**12**	23.657	32.000
Robust path, $\omega = .2$	70	**12**	1131.421	1136.000
Robust path, $\omega = .2$	150	8	1800.000	2416.000
Direct boundary, $\omega = .25$	150	16	2416.000	2416.000

The bounded-noise stress study uses $\epsilon = .02$, $\eta = .004$, $\delta = .05$, and 5,000 Bernoulli datasets per fixture. Two exact-model means and 36 distinct design-generated residual patterns yield 152 method–fixture cells for robust 16-, 12-, and 8-source plans and direct. Their largest simultaneous Monte Carlo upper failure bounds are 0.5176%, 0.3247%, 0.8299%, and 1.2993%, respectively, below 5%. Removing the residual allowance causes some compact-design intervals to fail in every repetition. These checks validate the implementation on supplied-allowance fixtures; proofs establish statements over the full model class.

D Structured-score experiments: protocol and uncertainty

D.1 Outcome construction and training

The saved implementation uses

$$Y(z, \xi) = .05 + \frac{.9}{h + |E|} \left(\sum_i R_i(a_i, \xi) + \sum_{ij \in E} R_{ij}(a_i, a_j, \xi) \right). \tag{28}$$

Here $a_i = \pi_i^{z_i}(\xi_i)$ and $R_i, R_{ij} \in \{0, 1\}$. Its range is $[.05, .95]$. Each role responds only to exogenous context, so node terms have degree at most one and pair terms degree at most two in binary policy choices. Shared exogenous randomness does not alter that degree.

Contexts contain an intercept, two shared standard-normal variables, two independent role-local normals, and a shared/local product. Population-specific response parameters are independently drawn. Each role has four actions and two variants. Each policy is trained on 4,096 randomized-action contextual-bandit examples with Bernoulli rewards. Variant reward-discount vectors are $(1, .9, .65, .35)$ and $(.5, .7, .95, 1)$. Ridge action-value fits use regularization three and choose the largest fitted action value. Qualification on 2,048 new contexts records local reward and disagreement without filtering out weak populations.

Node acceptance uses a population-specific logistic response and exogenous threshold. Resource-family pair scores reward distinct actions or joint action zero, subject to a pair-local capacity test. W6 instead requires that the sum of two action tags modulo four differs from three, together with a pair-local acceptance test. With probability .35, tests share a common threshold; otherwise thresholds are independent. The scalar acquisition interface hides primitive component values only within that declared access contract. Richer component/replay controls explicitly receive them.

P6 targets are a sorted 32-element sample, with catalogue seed 160922, from six-role vectors with more than two ones. The list is non-affine and outcome-independent. The other families use the full cube. Configuration IDs are lexicographic, first role most significant. Saved policies retain target vectors, feature matrices, learned weights, population parameters, qualification records, and seeds.

D.2 Development, confirmation, and reference separation

Development uses populations 910000–910002 and five repetitions. Feature-only predicted completion checkpoints are added before the confirmation freeze; every aggregate comparator receives the same grid. Confirmation uses population IDs 920000–920029 with 50 repetitions each. Seeds distinguish family, population, training, qualification, reference, method, and acquisition stage. Sensitivities at $.05$ and $.01$ reuse the frozen populations and references and therefore constitute paired conditions, not fresh training replications.

Primary acquisition caps are 131,072, 262,144, 1,048,576, and 262,144 for P6, B33, B35, and W6. Sensitivity caps scale by $.16$ and four and round up to a multiple of 256. Noncompleted repetitions receive their cap. Only consumed direct-stream prefixes count toward acquisition; potential simulator observations generated but not consumed are retained as backend overhead, not presented as avoided wall-clock work.

Independent references use continuous-context coefficient streams at eight predetermined checkpoints from 65,536 to 8,388,608 contexts. Target-wise empirical-Bernstein bounds are intersected until maximum reference half-width is at most $.001$. The reference error budget is $.001$ per family, split over 30 populations and eight looks; the four-family union budget is $.004$. Reference records contain chunk seeds, coefficient sums and cross-products.

Reference-confirmed, definite-miss, and indeterminate classifications follow Appendix B.3; they are conditional on reference validity and are not population coverage confidence intervals.

D.3 Plan pools, inference, and statistical units

P6 has 22 feature-only plans including direct, all-source, target-only, local controls, Walsh cosets, QR-derived supports, and coset unions. Other families retain their frozen pools. A common planning variance $.05$ ranks plans but is not a variance upper bound used in inference. The pool chooses the earliest predicted completion checkpoint, then predicted radius, support size, and a fixed name order. Every selected plan in the four primary families aliases the corresponding classical and finite-pool reference. Full-catalogue all-source and target-only sharing also alias.

Aggregate plans have at most twelve common fixed checkpoints with the error budget divided over the actual grid. Eight base looks are augmented by feature-only completion predictions. Charged pilots use 64 observations per source, exclude pilot outcomes from confirmation, and fix subsequent allocation conditional on the pilot. Direct-CS initially samples every target and then the largest current width in batches of 64, with first-index tie-breaking. Gamma-exponential tuning is fixed from the cap and catalogue size; predictable variance terms use previous observations only. Confidence intersections with $[0, 1]$ are allowed; inconsistency leads to abstention.

Each population contributes its average capped cost over 50 repetitions. Reported mean-cost intervals use t_{29} over the 30 independent population means. Paired log cost ratios give geometric-saving intervals; these differ from ratios of arithmetic mean costs. The primary direct comparisons' four-family adjusted intervals are $[25.87, 27.11]\%$, $[51.12, 52.92]\%$, $[81.19, 81.44]\%$, and $[55.86, 60.07]\%$. A fixed reporting grid can produce zero observed cost variance without implying universal determinism.

D.4 Acquisition costs and precision sensitivity

Table 6 summarizes acquisition costs for the four families. Its feature ranks are 12, 16, 24, and 13 for P6, B33, B35, and W6. Savings against target-only sharing are recomputed as $100(1 - \overline{C}_{\text{selected}}/\overline{C}_{\text{targets}})$ from the same stored means, not from rounded table entries. They are descriptive secondary contrasts, not a new confirmatory family.

The paired geometric-saving 95% intervals against Direct-CS are $[26.02, 26.97]\%$, $[51.33, 52.71]\%$, $[81.22, 81.41]\%$, and $[56.37, 59.61]\%$. W6 Direct-CS has one definite miss and one indeterminate repetition out of 1,500. Reference-confirmed coverage is an observed classification, not a population guarantee. Eight sensitivity conditions at .05 and .01 reuse the frozen populations. Savings increase over these three tested levels, not necessarily beyond them. At B35 precision .01, arithmetic mean costs 427,776 and 2,578,163.88 imply 83.41% savings. Figure 4 uses paired geometric summaries and matching four-family-adjusted intervals within each precision.

Table 6: **Structured acquisition at equal precision.** $\epsilon = .02$, $\delta = .05$, zero allowance/setup price; all methods complete 1,500/1,500 runs. $\pm$: population-t 95% half-width over 30 means; zero half-widths omitted. Savings use arithmetic means. Selected, classical, and pool-reference plans coincide; all selected runs are reference-confirmed.

Family	Targets / opens	Mean scalar acquisition (thousands)			Saving (%)	
		Direct-CS	Target sharing	**Selected / classical**	vs. CS	vs. sharing
P6	32/16	$78.72 \pm .51$	$65.56 \pm .02$	$\mathbf{57.86}$	26.51	11.75
B33	64/16	$166.92 \pm .77$	$104.22 \pm .73$	$\mathbf{80.15} \pm 1.44$	51.98	23.10
B35	256/32	729.95 ± 3.74	230.91	$\mathbf{136.38} \pm .39$	81.32	40.94
W6	64/16	171.86 ± 1.23	90.11	$\mathbf{72.61} \pm 3.24$	57.75	19.42

Every CACE sensitivity repetition contains its independent reference intervals. At .05 the direct procedures also have 100% reference-confirmed coverage. At .01, direct observed-coverage brackets are $[99.47, 100]\%$, $[99.80, 100]\%$, $[99.47, 99.93]\%$, and $[99.27, 100]\%$. B35 has one definite miss and seven indeterminate repetitions; P6, B33, and W6 have zero definite misses and eight, three, and eleven indeterminate repetitions. These are uncertainty brackets for classifying observed repetitions, not confidence intervals on a population rate.

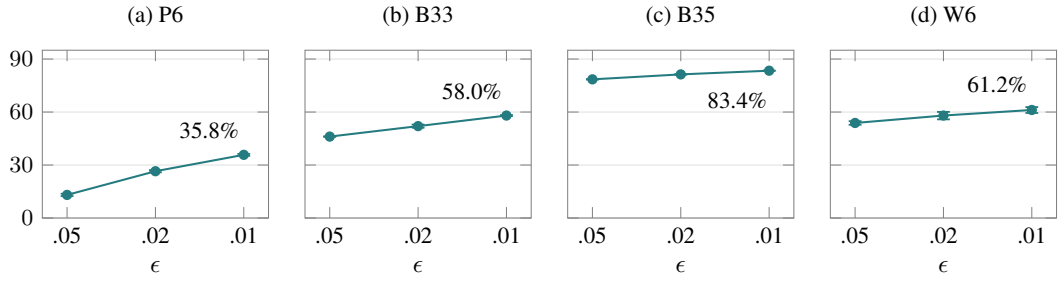

Figure 4: **Exact-model savings across tested precisions.** (a–d) Paired geometric savings against Direct-CS (%). Population-t intervals use four-family correction within each precision. All methods complete, with populations reused; allowance and setup price are zero. Bars may be narrower than markers.

Table 7: Measured precision sensitivities. Costs are mean consumed scalar observations, not thousands. Both CACE and Direct-CS complete in every sensitivity repetition. Setup price is zero, and the same 30 populations are reused at each precision.

ϵ	Family	CACE cost	Direct-CS cost	Saving (%)	Completion (%)
.05	P6	14,848.00	17,085.06	13.09	100
.05	B33	19,629.06	36,420.14	46.10	100
.05	B35	34,624.34	161,110.23	78.51	100
.05	W6	17,289.56	37,384.28	53.75	100
.01	P6	180,480.00	281,065.05	35.79	100
.01	B33	249,330.01	593,805.53	58.01	100
.01	B35	427,776.00	2,578,163.88	83.41	100
.01	W6	238,740.82	612,090.03	61.00	100

Table 8: Measured population-level sensitivity uncertainty. Intervals use log ratios of population mean costs and t_{29} quantiles. Bonferroni adjustment applies over four families separately at each precision, not jointly over every family–precision condition.

ϵ	Family	Geometric saving (%)	95% interval	Four-family interval
.05	P6	13.08	$[12.57, 13.60]$	$[12.41, 13.75]$
.05	B33	46.10	$[45.97, 46.24]$	$[45.93, 46.28]$
.05	B35	78.51	$[78.43, 78.59]$	$[78.40, 78.61]$
.05	W6	53.81	$[53.03, 54.59]$	$[52.78, 54.82]$
.01	P6	35.78	$[35.35, 36.20]$	$[35.22, 36.33]$
.01	B33	58.02	$[57.79, 58.24]$	$[57.72, 58.31]$
.01	B35	83.41	$[83.32, 83.50]$	$[83.29, 83.52]$
.01	W6	61.17	$[59.87, 62.44]$	$[59.46, 62.81]$

D.5 Recorded stress, richer access, and reuse

The saved policy stress adds $\eta s(z)$ with a worst-sign pattern fixed from design weights before outcomes. At allowance ratios $0, .05, .10, .15, .20, .25$, protected and noise-only intervals use identical counts and weights. At every positive ratio, protected intervals cover the references but do not reach the requested width at those unchanged counts. At ratio .25, P6 noise-only coverage is bracketed by $[20.40, 36.27]\%$. No reoptimized protected completion cost is inferred from this diagnostic.

W6 component/replay caches both node variants and all four pair variants. It costs $2h + 4|E| = 36$ primitive tests per context, versus $h + |E| = 12$ per scalar score, or three scalar-equivalent units. Its all-target bounds use a union bound and permit dependence across replayed target scores. A separate deterministic 512-context library can be fully enumerated at 16,384 calls for P6, 32,768 for B33/W6, and 131,072 for B35. Those exact finite-library means are different estimands from the continuous task law.

For reuse, 200 feasible-subset/known-cost queries per population are applied to each of its 50 interval catalogues. Targets are included with probability .4, with a nonempty subset guaranteed; costs are uniform on $[0, .05]$. Methods reuse cached evidence without new calls. Midpoint-diagnostic CACE regret means range from .000172 to .000201; the largest reference-based regret upper bound is .024525, below the conditional $2\epsilon = .04$ theorem bound. Midpoint references are diagnostics, not exact truth.

D.6 Planning, access, and precision diagnostics

Removing P6's cell-variance planning penalty selects 64 rather than 16 sources and increases acquisition from 57.856k to 77.824k with final inference unchanged. A charged independent compact pilot uses 65.950k. The resulting comparison isolates planning, not a different final confidence rule. At unchanged nominal counts, supplying a positive residual allowance can retain observed

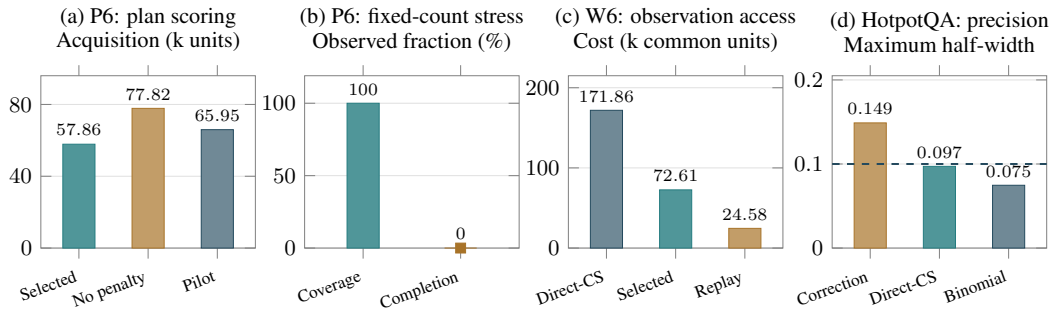

Figure 5: **Measured planning, access, and precision diagnostics.** (a) Removing P6's cell-variance planning penalty changes the selected support and raises acquisition cost. (b) At $\eta/\epsilon = .25$, unchanged-count protected intervals are reference-confirmed but do not complete. (c) W6 replay uses richer component access, with costs converted to common units. (d) Correction misses the .10 precision target, while both direct controls reach it. Uncertainty and accounting are in Appendix D.

coverage while preventing completed precision. W6 component/replay acquisition instead changes the observation interface: it completes with 24.576k scalar-equivalent units versus 72.609k for selected scalar acquisition and 171.858k for Direct-CS. It is not an equal-access comparison.

Panel (d) summarizes the external comparison detailed in Appendix G.

E Operational cost experiments and sensitivity sweeps

The operational study uses fixed-Hoeffding and EB procedures on five primary settings, with 30 populations and 50 nested acquisitions per population. It uses populations distinct from the nominal study; supporting precision, boundary, and price conditions reuse the corresponding operational populations. Complete conditions and descriptive uncertainty appear in Appendix E.3.

Two distinct engines. Both confidence engines are complete throughout the campaign. The primary arms use fixed-Hoeffding and aggregate empirical-Bernstein inference on the five P_4, $K_{3,3}$, and $K_{3,5}$ settings. Matched EB methods share the reporting grid, error split, caps, and cell-variance penalty; the per-look error allocation is $\delta/64$. The Hoeffding arm uses its fixed certificate. Direct-CS has one charged, excluded pilot per target to initialize the predictable mean, followed by width-adaptive batches of 64. This makes replay under deterministic target shifts translation equivariant. Its comparator is not conflated with Direct-EB in Table 1. The operational endpoint uses untruncated intervals; range clipping cannot produce apparent completion.

Settings and population separation. Confirmation uses population IDs 1920000–1920029 and seed prefix $(20260919, 19041)$, separate from the earlier campaign. Each policy retains the 4,096-example training and 2,048-context qualification protocol, with no population exclusion. The same trained population is reused across its conditions and engines, giving paired conditions rather than additional training replications. The five primary (ω, t) settings are $(0, 1), (.2, 150)$ for P_4, $(.1, 0)$ for $K_{3,3}$, and $(.2, 0), (.4, 0)$ for $K_{3,5}$, all at $\epsilon = .02$. Assigned opening price is $\kappa/c = tL$, with $L = \log(2M/\delta)/(2\epsilon^2)$. At high-setup P_4 this is approximately 1,211,525.28 scalar-call cost units per opened source.

Residual fixtures comprise an outcome-free union of sign patterns generated from all candidate designs, plus the exact-model fixture. Every method receives the same fixture at a given condition. With $Y_\eta = Y + \eta s(z)$, the supplied construction and the bounded range justify η; it is not estimated from observed residuals. Acquisition uses independent base scalar scores, then replays the frozen deterministic shifts with common random numbers. Fixtures are dependent views of each repetition and do not enlarge the independent-population sample size.

Controls and cost ledger. The frozen pool contains direct, protected nominal/classical plans, robust compact path alternatives, and full-support reoptimized candidates. The full-support conic candidate

uses XOR symmetry and uniform counts, with a count-search tolerance of one observation per source; this does not claim global arbitrary-allocation or support optimality under EB. The common planning variance is .05, independent of reference outcomes. Infeasible fixed estimators are rejected before acquisition, and exact plan aliases share evidence. The measured controls and their rejection statuses are in Table 9.

The reported charge is $\kappa s + cN$, including consumed scalar observations and excluded pilots. Pre-sampling screening and supplied simulator assurance use no acquisition calls; their computational time is not measured as deployment cost. Caps derive from the matched worst-case direct acquisition plus direct opening charges, and capped failures are retained. Training, reference simulation, and unconsumed backend draws are recorded separately. No acquisition weights or supports were updated after seeing confirmation outcomes. The high-setup P_4 selected charge is 13,803,178.26 cost units versus 19,476,359.29 for Direct-EB, despite using more scalar observations; the advantage depends on the assigned opening price.

Analysis and completion criteria. The primary endpoint is capped total cost to all-target half-width ϵ, with completion and coverage reported alongside it. Each population supplies a mean over its 50 repetitions. Paired log ratios and t_{29} intervals summarize population-level effects, with the ten-contrast correction in Appendix E.1. Ratios of arithmetic means are reported separately. No target, fixture, or nested repetition is treated as an independent trained system, and ties from exact aliases are not converted into small wins.

E.1 Primary operational contrasts and measured controls

The five primary settings in Table 1 were specified before confirmation; the wider grid is descriptive.

The graph, admissible sources, assurance construction, support pool, direct certificate, reporting schedule, cap, and contrast family were frozen before confirmation. Development used a separate namespace. The ten primary contrast slots compare selected with matched direct and with protected nominal in each of five EB settings. The paired log-ratio intervals use Bonferroni level $.05/10$ without recycling slots for rejected estimators. The full allowance–price grid and Direct-CS comparisons remain descriptive.

The selected feature-only support is fixed across populations within each condition. Comparisons use the 30 population means defined in Appendix B.3; the 50 nested repetitions do not constitute additional training replications.

Geometric paired ratios and ratios of arithmetic means are distinct summaries. All measured selected, nominal, matched-direct, and Direct-CS arms in the primary study complete in every repetition; pre-sampling rejections are excluded from that statement. For the estimable selected–baseline comparisons, observed paired completion differences and their population-t intervals are 0 and $[0, 0]$. This reflects zero observed variation, not proof of a zero population failure probability. Cost is not conditioned on successful runs, and unknown reference classifications remain indeterminate.

A plan rejected by the outcome-free screen has no acquired samples and no measured coverage or stopping cost. The high-setup P_4 and both $K_{3,5}$ protected nominal estimators are rejected because their fixed residual floor reaches the width budget. This is not a zero-cost completed procedure, nor a claim that every weight choice on a full support is infeasible. The low-setup P_4 selected and nominal plans share their evidence; high-error $K_{3,5}$ selected and direct likewise share evidence and are reported as an exact tie.

In the EB mild-error $K_{3,3}$ condition, the selected plan is the reoptimized full-support conic candidate. At moderate-error $K_{3,5}$ it also aliases reoptimized full support. Their arithmetic savings against Direct-CS are only 12.59% and 2.95%, respectively, compared with 60.55% and 47.98% against matched Direct-EB. At high-error $K_{3,5}$, the selected/direct-EB ratio is one, whereas selected/Direct-CS is 1.8256. These engine-dependent outcomes are retained rather than interpreted as universal direct-method dominance.

References and post-hoc precision refinement. The original operational references use eight fixed looks from 65,536 to 8,388,608 contexts, target half-width .001, and budget $.001/(4 \cdot 30)$ per

Table 9: Operational controls. Point ratios compare arithmetic mean costs. Protected-nominal intervals use paired geometric ratios with the same ten-contrast adjustment as Table 13. "Rejected" means the fixed estimator exhausts the residual-width budget before sampling; no acquisition ratio is assigned. Direct-CS ratios are descriptive additional comparisons. A dash denotes no opened nominal support after pre-sampling rejection. All measured arms shown complete in every repetition. The B33 contrast changes support as well as weights.

Family, (ω, t)	Opens nominal / selected	Selected/ nominal	Adjusted geometric ratio interval	Selected/ Direct-CS
$P_4, 0, 1$	8 / 8	1.0000	$[1.0000, 1.0000]$	0.7555
$P_4, .2, 150$	– / 8	**Rejected**	–	0.7102
$K_{3,3}, .1, 0$	16 / 64	0.7379	$[0.7276, 0.7488]$	0.8741
$K_{3,5}, .2, 0$	– / 256	**Rejected**	–	0.9705
$K_{3,5}, .4, 0$	– / 256	**Rejected**	–	1.8256

Table 10: Completed primary fixed-Hoeffding arm on the same fresh populations. The denominator is its own matched direct certificate, not Direct-EB or Direct-CS. P4 coverage incorporates the same post-hoc reference refinement. These rows do not replace the separate boundary and price sweeps.

Family, (ω, t)	Opens	Cost/direct	Completion	Coverage bracket (%)
$P_4, 0, 1$	12	0.7393	1500/1500	[100.00,100.00]
$P_4, .2, 150$	8	0.7450	1500/1500	[100.00,100.00]
$K_{3,3}, .1, 0$	64	0.4484	1500/1500	[100.00,100.00]
$K_{3,5}, .2, 0$	256	0.5000	1500/1500	[100.00,100.00]
$K_{3,5}, .4, 0$	256	1.0000	1500/1500	[100.00,100.00]

population, reserving a total .001 for the four-family operational campaign. The original high-setup P_4 EB worst-fixture counts were 1,155 confirmed, zero definite misses, and 345 indeterminate.

After inspecting that ambiguity, a distinct protocol was fixed for fresh P4 references: seed prefix $(20260920, 441001)$, requested half-width .00025, eight looks from 2^{20} through 2^{27}, and an additional total reference budget .001 split over 30 populations and all looks and targets. Stopping depended only on reference precision, not on acquisition coverage. Every population stopped at 2^{25} contexts, totaling 1,006,632,960; maximum target half-widths ranged from .000230473 to .000243569. Original reference and acquisition records were preserved, and the fresh bounds alone supply the refined P4 classification.

The refined high-setup P4 EB classification has 1,499 jointly confirmed repetitions, zero definite misses, and one indeterminate repetition across 117 frozen fixtures. The other three primary P4 condition–engine combinations have 1,500 jointly confirmed repetitions. This post-hoc reference refinement changes neither the acquisition observations nor the selected estimators.

Joint reliance on the original operational references and the refinement has a conservative reference failure bound .002. Including the earlier nominal campaign's .004 reference budget gives .006 for a joint reference event across both campaigns and this refinement. These reference budgets are distinct from each acquisition procedure's $\delta = .05$, and the observed coverage brackets are not confidence intervals for population coverage probability.

Completed supporting settings. The original operational sensitivities at $\epsilon = .01, .05$ reuse the primary populations and fix normalized (ω, t). Consequently $\eta = \omega\epsilon$ and $\kappa/c = t\log(2M/\delta)/(2\epsilon^2)$ change with precision. For high-setup P4, $(\eta, \kappa/c)$ is $(.002, 4{,}846{,}101.13)$, $(.004, 1{,}211{,}525.28)$, and $(.010, 193{,}844.05)$ at $\epsilon = .01, .02, .05$. The distinct fixed-allowance, fixed-price study is in Appendix F. The boundary grid includes $0, .05, .10, .20, .25, 1/3, .40$ and neighbors $\omega_* \pm .001$ of the applicable threshold for P_4, $K_{1,3}$, $K_{3,3}$, and $K_{3,5}$, at $t = 0$ and $\epsilon = .02$. The P4 price grid is $t \in \{0, .5, .9, 1, 1.1, 10, 69, 69.02, 77.14, 77.15, 100, 150\}$ at $\omega = 0, .2$ and $\epsilon = .02$. Duplicate grid points are removed. This yields 10 precision, 36 boundary, and 24 price conditions per engine. The boundary-only star family has its own 30 fresh populations; other conditions reuse the corresponding

Table 11: **Post-hoc reuse of existing finer P4 references.** All rows use $\omega = .2$ and the selected method; H denotes Hoeffding. Brackets are the worst-fixture observed classifications. Counts in the final column are joint across all frozen fixtures in that condition. Saved policies match exactly across studies; acquisition intervals and costs are unchanged. The primary refinement used fresh reference samples; the subsequent supporting-sweep reclassification used no additional samples. Distinct conditions reuse populations, not necessarily identical acquisition draws. These brackets are not confidence intervals for future coverage probability.

Study, ϵ, t, engine	Original (%)	Refined (%)	Confirmed/missed/unknown
Primary, $0.02, 150$, EB	[77.00,100.00]	[99.93,100.00]	1499/0/1
Primary, $0.02, 150$, H	[99.93,100.00]	[100.00,100.00]	1500/0/0
Precision, $0.01, 150$, EB	[0.20,100.00]	[98.60,100.00]	1479/0/21
Precision, $0.01, 150$, H	[62.13,100.00]	[100.00,100.00]	1500/0/0
Price, $0.02, 100$, EB	[89.60,100.00]	[100.00,100.00]	1500/0/0
Price, $0.02, 150$, EB	[76.53,100.00]	[100.00,100.00]	1500/0/0
Price, $0.02, 69.02$, EB	[97.27,100.00]	[100.00,100.00]	1500/0/0
Price, $0.02, 69$, EB	[96.07,100.00]	[100.00,100.00]	1500/0/0
Price, $0.02, 77.14$, EB	[94.73,100.00]	[100.00,100.00]	1500/0/0
Price, $0.02, 77.15$, EB	[95.47,100.00]	[100.00,100.00]	1500/0/0

frozen primary policies. The older nominal sensitivities in Table 7 remain a separate completed campaign.

Post-hoc reuse of verified references. After the supporting studies, the same finer P4 reference intervals were reused for the ten selected-method cases with unresolved original classifications, including the two primary cases. The populations and frozen policies match; no new reference samples, acquisition changes, or training replications were introduced. Table 11 retains the original and refined brackets. Reusing the same simultaneous event leaves the operational reference failure bound at .002.

E.2 Cost summaries and sweep interpretation

At normalized precision .01, B33 at $\omega = .1$ and B35 at $\omega = .2$ save 28.92% and 21.69% against Direct-CS; at .05 they cost 17.20% and 28.23% more. Among the 36 EB boundary conditions, nine beat Direct-CS and 27 cost more. These paired, descriptive sensitivities fix (ω, t) and therefore change absolute allowance and positive setup price with precision; the fixed-physical controls are in Appendix F.

Of the 150 condition–engine cases, 148 have reference brackets $[100, 100]\%$. High-setup P4 under EB retains $[99.93, 100]\%$ at primary precision and $[98.60, 100]\%$ at normalized precision .01, with one and 21 indeterminate repetitions and no definite misses. These classifications incorporate the disclosed finer-reference reuse.

Table 12: **Completed operational sweeps.** Records include both engines and 30 populations per condition; each record has 50 nested repetitions. Ratios compare arithmetic mean charged costs of the selected EB method with Direct-EB or Direct-CS. Every selected repetition completes. These counts describe a frozen, paired grid, not independent wins or new confirmatory contrasts. Detailed values and descriptive paired uncertainty appear in Appendix E.3.

Sweep	Cases/engine	Records	$R_{\mathrm{EB}} < 1 / = 1$	$R_{\mathrm{CS}} < 1 / > 1$
Precision	10	600	8 / 2	6 / 4
Boundary	36	2,160	20 / 16	9 / 27
Price	24	1,440	24 / 0	18 / 6

Table 13: **Paired operational cost uncertainty.** Savings use ratios of arithmetic mean costs relative to matched Direct-EB. Geometric ratios and their intervals use the 30 population log ratios and t_{29} quantiles. Each interval uses $\alpha = .05/10$ for the prespecified ten-contrast family (selected versus direct and protected nominal in five EB settings); infeasible contrasts are not reassigned that budget. The tinted high-error tie is an exact evidence alias. All four nonzero matched-EB reductions survive this adjustment.

Family, (ω, t)	Saving (%)	Geometric ratio	Adjusted ratio interval
$P_4, 0, 1$	39.90	0.6008	$[0.5937, 0.6080]$
$P_4, .2, 150$	29.13	0.7087	$[0.7087, 0.7088]$
$K_{3,3}, .1, 0$	60.55	0.3945	$[0.3909, 0.3982]$
$K_{3,5}, .2, 0$	47.98	0.5206	$[0.5094, 0.5321]$
$K_{3,5}, .4, 0$	0.00	**1.0000**	$[1.0000, 1.0000]$

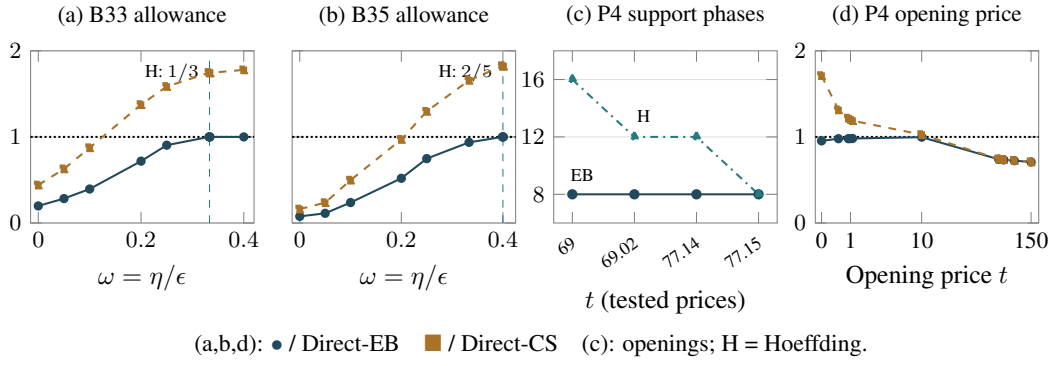

Figure 6: **Measured allowance and price regimes.** (a,b) Arithmetic cost ratios for nine EB allowance cases per family, $\epsilon = .02$, $t = 0$; vertical lines are Hoeffding boundaries. (c) Four price checks reproduce H's $16 \to 12 \to 8$ support phases. (d) Twelve P4 prices at $\omega = .2$, shown on a $\log(1 + t)$ axis with original-price labels. Below one favors sharing; uncertainty is in Appendix E.3.

E.3 Complete sweep tables and descriptive uncertainty

Tables 14–21 report every supporting condition in both engines. S is the selected number of opened sources. R_D is the ratio of arithmetic mean charged costs to the matched direct engine; R_{CS} uses Direct-CS. G_{CS} is the geometric mean of 30 paired population cost ratios, with a descriptive 95% interval from the population log ratios and t_{29}. Arithmetic and geometric ratios are different summaries. These intervals are not adjusted across the supporting grid; the ten-contrast confirmatory correction applies only to the primary EB family. Every selected, matched-direct, and Direct-CS arm in these sweeps completes all 1,500 repetitions per condition. Selected observed coverage is fully reference-confirmed in the boundary and price grids after the disclosed P4 reference reuse. Tight-precision, high-setup P4 under EB retains the $[98.60, 100]\%$ bracket in Table 11; other selected precision conditions are fully reference-confirmed. The complete arithmetic ratios, costs, scalar calls and original/refined coverage brackets are also recorded in the separate reproducibility archive.

Table 14: Completed operational precision sweep, aggregate EB. Each row uses the same 30 frozen populations as its primary setting; precision changes the assigned opening price through $\kappa/c = t\log(2M/\delta)/(2\epsilon^2)$. Every displayed arm completes; the additional conditions are paired, not fresh training replications.

Family, (ω, t)	ϵ	S	R_D	R_{CS}	G_{CS} [95% interval]
P_4, $(0, 1)$	0.01	8	0.5862	0.7030	0.7029 [0.6986, 0.7072]
P_4, $(0, 1)$	0.05	8	0.5974	0.8678	0.8677 [0.8645, 0.8710]
P_4, $(0.2, 150)$	0.01	8	0.7050	0.7060	0.7056 [0.6972, 0.7142]
P_4, $(0.2, 150)$	0.05	8	0.7085	0.7113	0.7113 [0.7113, 0.7114]
$K_{3,3}$, $(0.1, 0)$	0.01	64	0.4370	0.7108	0.7108 [0.7092, 0.7124]
$K_{3,3}$, $(0.1, 0)$	0.05	64	0.4300	1.1720	1.1720 [1.1697, 1.1743]
$K_{3,5}$, $(0.2, 0)$	0.01	256	0.4926	0.7831	0.7831 [0.7812, 0.7851]
$K_{3,5}$, $(0.2, 0)$	0.05	256	0.5531	1.2823	1.2823 [1.2794, 1.2851]
$K_{3,5}$, $(0.4, 0)$	0.01	256	1.0000	1.5562	1.5551 [1.5337, 1.5767]
$K_{3,5}$, $(0.4, 0)$	0.05	256	1.0000	2.3052	2.3049 [2.2933, 2.3166]

Table 15: Completed operational precision sweep, fixed Hoeffding. Each row uses the same 30 frozen populations as its primary setting; precision changes the assigned opening price through $\kappa/c = t\log(2M/\delta)/(2\epsilon^2)$. Every displayed arm completes; the additional conditions are paired, not fresh training replications.

Family, (ω, t)	ϵ	S	R_D	R_{CS}	G_{CS} [95% interval]
P_4, $(0, 1)$	0.01	12	0.7393	1.1216	1.1216 [1.1203, 1.1230]
P_4, $(0, 1)$	0.05	12	0.7393	1.0260	1.0260 [1.0246, 1.0274]
P_4, $(0.2, 150)$	0.01	8	0.7450	0.7484	0.7484 [0.7484, 0.7484]
P_4, $(0.2, 150)$	0.05	8	0.7450	0.7478	0.7478 [0.7478, 0.7478]
$K_{3,3}$, $(0.1, 0)$	0.01	64	0.4484	1.5866	1.5867 [1.5822, 1.5911]
$K_{3,3}$, $(0.1, 0)$	0.05	64	0.4484	1.1359	1.1359 [1.1330, 1.1388]
$K_{3,5}$, $(0.2, 0)$	0.01	256	0.5000	1.8807	1.8808 [1.8752, 1.8863]
$K_{3,5}$, $(0.2, 0)$	0.05	256	0.5003	1.3412	1.3412 [1.3377, 1.3447]
$K_{3,5}$, $(0.4, 0)$	0.01	256	1.0000	3.7611	3.7612 [3.7501, 3.7723]
$K_{3,5}$, $(0.4, 0)$	0.05	256	1.0000	2.6812	2.6813 [2.6743, 2.6882]

Table 16: Completed P_4 boundary sweep at $t = 0$, $\epsilon = .02$; H denotes fixed Hoeffding. Tint marks the analytic boundary. All selected repetitions complete and are reference-confirmed. Ratios of one to the matched engine are retained as direct-plan aliases. Rounded or finite-pool choices just below a theoretical threshold do not identify a new exact boundary.

ω	S_{EB}	$R_{D,\mathrm{EB}}$	$G_{\mathrm{CS,EB}}$ [95%]	S_{H}	$R_{D,\mathrm{H}}$	$G_{\mathrm{CS,H}}$ [95%]
0	16	0.5077	0.9675 [0.9642, 0.9709]	16	0.5001	1.4435 [1.4365, 1.4505]
0.05	16	0.5365	1.1434 [1.1381, 1.1488]	16	0.6098	1.7603 [1.7516, 1.7690]
0.1	16	0.6286	1.3390 [1.3332, 1.3449]	16	0.7353	2.1227 [2.1124, 2.1331]
0.2	16	0.9532	1.7119 [1.7057, 1.7182]	16	0.9616	2.7751 [2.7616, 2.7886]
0.249	16	1.0000	1.7701 [1.7614, 1.7788]	16	1.0000	2.8865 [2.8724, 2.9006]
0.25	16	1.0000	1.7709 [1.7619, 1.7799]	16	1.0000	2.8874 [2.8731, 2.9018]
0.251	16	1.0000	1.7707 [1.7607, 1.7808]	16	1.0000	2.8866 [2.8728, 2.9006]
0.333333	16	1.0000	1.7738 [1.7673, 1.7803]	16	1.0000	2.8877 [2.8737, 2.9017]
0.4	16	1.0000	1.7853 [1.7791, 1.7915]	16	1.0000	2.8857 [2.8718, 2.8996]

Table 17: Completed $K_{1,3}$ boundary sweep at $t = 0$, $\epsilon = .02$; H denotes fixed Hoeffding. Tint marks the analytic boundary. All selected repetitions complete and are reference-confirmed. Ratios of one to the matched engine are retained as direct-plan aliases. Rounded or finite-pool choices just below a theoretical threshold do not identify a new exact boundary.

ω	S_{EB}	$R_{D,\mathrm{EB}}$	$G_{\mathrm{CS,EB}}$ [95%]	S_{H}	$R_{D,\mathrm{H}}$	$G_{\mathrm{CS,H}}$ [95%]
0	8	0.4241	0.9019 [0.8994, 0.9043]	8	0.5000	1.4398 [1.4349, 1.4446]
0.05	8	0.4987	1.0609 [1.0544, 1.0673]	16	0.5807	1.6723 [1.6670, 1.6777]
0.1	16	0.5881	1.2505 [1.2445, 1.2565]	16	0.6712	1.9326 [1.9262, 1.9390]
0.2	16	0.7292	1.5510 [1.5408, 1.5613]	16	0.8621	2.4817 [2.4736, 2.4900]
0.25	16	0.9367	1.7020 [1.6886, 1.7156]	16	0.9412	2.7102 [2.7009, 2.7195]
0.332333	16	1.0000	1.7873 [1.7779, 1.7968]	16	1.0000	2.8802 [2.8707, 2.8898]
0.333333	16	1.0000	1.7872 [1.7774, 1.7970]	16	1.0000	2.8798 [2.8700, 2.8896]
0.334333	16	1.0000	1.7865 [1.7761, 1.7969]	16	1.0000	2.8786 [2.8692, 2.8880]
0.4	16	1.0000	1.7883 [1.7782, 1.7985]	16	1.0000	2.8792 [2.8699, 2.8885]

Table 18: Completed $K_{3,3}$ boundary sweep at $t = 0$, $\epsilon = .02$; H denotes fixed Hoeffding. Tint marks the analytic boundary. All selected repetitions complete and are reference-confirmed. Ratios of one to the matched engine are retained as direct-plan aliases. Rounded or finite-pool choices just below a theoretical threshold do not identify a new exact boundary.

ω	S_{EB}	$R_{D,\mathrm{EB}}$	$G_{\mathrm{CS,EB}}$ [95%]	S_{H}	$R_{D,\mathrm{H}}$	$G_{\mathrm{CS,H}}$ [95%]
0	16	0.1984	0.4407 [0.4351, 0.4463]	16	0.2500	0.8110 [0.8087, 0.8132]
0.05	16	0.2839	0.6286 [0.6271, 0.6300]	64	0.3325	1.0787 [1.0757, 1.0818]
0.1	64	0.3950	0.8742 [0.8723, 0.8761]	64	0.4484	1.4549 [1.4509, 1.4589]
0.2	64	0.7186	1.3747 [1.3652, 1.3843]	64	0.7576	2.4577 [2.4509, 2.4645]
0.25	64	0.9041	1.5850 [1.5733, 1.5969]	64	0.8889	2.8839 [2.8760, 2.8918]
0.332333	64	1.0000	1.7442 [1.7369, 1.7516]	64	1.0000	3.2442 [3.2350, 3.2533]
0.333333	64	1.0000	1.7452 [1.7384, 1.7520]	64	1.0000	3.2442 [3.2350, 3.2534]
0.334333	64	1.0000	1.7442 [1.7367, 1.7517]	64	1.0000	3.2441 [3.2351, 3.2532]
0.4	64	1.0000	1.7785 [1.7739, 1.7831]	64	1.0000	3.2442 [3.2352, 3.2532]

Table 19: Completed $K_{3,5}$ boundary sweep at $t = 0$, $\epsilon = .02$; H denotes fixed Hoeffding. Tint marks the analytic boundary. All selected repetitions complete and are reference-confirmed. Ratios of one to the matched engine are retained as direct-plan aliases. Rounded or finite-pool choices just below a theoretical threshold do not identify a new exact boundary.

ω	S_{EB}	$R_{D,\mathrm{EB}}$	$G_{\mathrm{CS,EB}}$ [95%]	S_{H}	$R_{D,\mathrm{H}}$	$G_{\mathrm{CS,H}}$ [95%]
0	32	0.0770	0.1619 [0.1616, 0.1623]	32	0.0937	0.3227 [0.3217, 0.3236]
0.05	32	0.1128	0.2361 [0.2355, 0.2367]	256	0.1381	0.4753 [0.4739, 0.4767]
0.1	32	0.2377	0.4979 [0.4967, 0.4991]	256	0.2128	0.7323 [0.7302, 0.7345]
0.2	256	0.5196	0.9706 [0.9682, 0.9730]	256	0.5000	1.7210 [1.7161, 1.7260]
0.25	256	0.7481	1.2961 [1.2790, 1.3133]	256	0.6957	2.3944 [2.3875, 2.4014]
0.333333	256	0.9370	1.6569 [1.6454, 1.6685]	256	0.9310	3.2047 [3.1954, 3.2140]
0.399	256	1.0000	1.8235 [1.8063, 1.8409]	256	1.0000	3.4418 [3.4319, 3.4518]
0.4	256	1.0000	1.8215 [1.8028, 1.8404]	256	1.0000	3.4419 [3.4319, 3.4519]
0.401	256	1.0000	1.8218 [1.8030, 1.8407]	256	1.0000	3.4420 [3.4320, 3.4520]

Table 20: Completed P_4 price sweep at $\omega = 0$, $\epsilon = .02$; H denotes fixed Hoeffding. All selected arms complete and have fully reference-confirmed observed coverage after the disclosed refinement reuse. t assigns an opening charge, not measured runtime; tint/bold mark the intermediate H support. Each engine retains its own selected support and matched direct denominator.

t	S_{EB}	$R_{D,\mathrm{EB}}$	$G_{\mathrm{CS,EB}}$ [95%]	S_{H}	$R_{D,\mathrm{H}}$	$G_{\mathrm{CS,H}}$ [95%]
0	16	0.5077	$0.9675\,[0.9642, 0.9709]$	16	0.5001	$1.4435\,[1.4365, 1.4505]$
0.5	8	0.6638	$0.8989\,[0.8913, 0.9066]$	16	0.6667	$1.1815\,[1.1792, 1.1839]$
0.9	8	0.6069	$0.7748\,[0.7690, 0.7806]$	16	0.7369	$1.1233\,[1.1218, 1.1248]$
1	8	0.6012	$0.7556\,[0.7504, 0.7609]$	**12**	0.7393	$1.0981\,[1.0967, 1.0995]$
1.1	8	0.5958	$0.7386\,[0.7339, 0.7434]$	8	0.7381	$1.0716\,[1.0703, 1.0728]$
10	8	0.5166	$0.5344\,[0.5337, 0.5351]$	8	0.5455	$0.5799\,[0.5798, 0.5800]$
69	8	0.5026	$0.5052\,[0.5050, 0.5053]$	8	0.5071	$0.5119\,[0.5119, 0.5119]$
69.02	8	0.5025	$0.5051\,[0.5050, 0.5053]$	8	0.5071	$0.5119\,[0.5119, 0.5119]$
77.14	8	0.5023	$0.5046\,[0.5045, 0.5047]$	8	0.5064	$0.5107\,[0.5107, 0.5107]$
77.15	8	0.5023	$0.5046\,[0.5045, 0.5047]$	8	0.5064	$0.5107\,[0.5107, 0.5107]$
100	8	0.5016	$0.5034\,[0.5034, 0.5035]$	8	0.5050	$0.5082\,[0.5082, 0.5082]$
150	8	0.5012	$0.5024\,[0.5023, 0.5024]$	8	0.5033	$0.5055\,[0.5055, 0.5055]$

Table 21: Completed P_4 price sweep at $\omega = 0.2$, $\epsilon = .02$; H denotes fixed Hoeffding. All selected arms complete and have fully reference-confirmed observed coverage after the disclosed refinement reuse. t assigns an opening charge, not measured runtime; tint/bold mark the intermediate H support. Each engine retains its own selected support and matched direct denominator.

t	S_{EB}	$R_{D,\mathrm{EB}}$	$G_{\mathrm{CS,EB}}$ [95%]	S_{H}	$R_{D,\mathrm{H}}$	$G_{\mathrm{CS,H}}$ [95%]
0	16	0.9542	$1.7120\,[1.7061, 1.7180]$	16	0.9616	$2.7758\,[2.7625, 2.7892]$
0.5	16	0.9794	$1.3138\,[1.3108, 1.3167]$	16	0.9744	$1.7270\,[1.7235, 1.7305]$
0.9	16	0.9796	$1.2176\,[1.2153, 1.2199]$	16	0.9798	$1.4935\,[1.4915, 1.4955]$
1	16	0.9809	$1.2022\,[1.2002, 1.2041]$	16	0.9808	$1.4569\,[1.4550, 1.4587]$
1.1	16	0.9817	$1.1884\,[1.1865, 1.1904]$	16	0.9817	$1.4254\,[1.4237, 1.4271]$
10	16	0.9970	$1.0271\,[1.0269, 1.0274]$	16	0.9965	$1.0595\,[1.0593, 1.0596]$
69	8	0.7394	$0.7427\,[0.7427, 0.7428]$	16	0.9995	$1.0089\,[1.0088, 1.0089]$
69.02	8	0.7394	$0.7427\,[0.7427, 0.7427]$	**12**	0.9994	$1.0088\,[1.0088, 1.0089]$
77.14	8	0.7347	$0.7377\,[0.7377, 0.7377]$	**12**	0.9735	$0.9817\,[0.9817, 0.9817]$
77.15	8	0.7347	$0.7377\,[0.7377, 0.7377]$	8	0.9734	$0.9817\,[0.9816, 0.9817]$
100	8	0.7243	$0.7265\,[0.7265, 0.7265]$	8	0.8663	$0.8720\,[0.8720, 0.8720]$
150	8	0.7087	$0.7102\,[0.7102, 0.7102]$	8	0.7450	$0.7483\,[0.7483, 0.7483]$

F SUPPORT, WEIGHTING, AND FIXED-PHYSICAL CONTROLS

The fixed-physical study holds allowance and opening price fixed while changing ϵ; the support/weight ablation changes the B33 estimator. Both use the same 30 operational populations with 50 nested repetitions. The two .02 anchors and matching-precision Direct-CS endpoints are reused, so they add no independent training populations. New acquisition streams append $(20260921, 7301)$ to distinguish them from the original campaign. Candidate weights were fixed before confirmation, and these comparisons do not change the primary contrast family.

Table 22: **Fixed-allowance, fixed-price sensitivity.** B33 holds $\eta = .002$, $\kappa/c = 0$; P4 holds $\eta = .004$, $\kappa/c = 1{,}211{,}525.28$. Tint and stars mark reused primary anchors; other selected/Direct-EB outcomes are new. All Direct-CS outcomes are previously completed matching-epsilon runs, reclassified at the new allowance and repriced from consumed calls. All displayed methods complete. Ratios use arithmetic mean costs; comparisons are descriptive.

Family	ϵ	ω	t	Opens	Selected cost / direct cost Direct-EB	Direct-CS
$K_{3,3}$	0.01	0.20	0	64	0.7008	1.1404
$K_{3,3}$	0.02*	0.10	0	64	0.3945	0.8741
$K_{3,3}$	0.05	0.04	0	16	0.2436	0.6621
P_4	0.01	0.40	37.5	16	1.0000	1.0052
P_4	0.02*	0.20	150	8	0.7087	0.7102
P_4	0.05	0.08	937.5	8	0.5013	0.5017

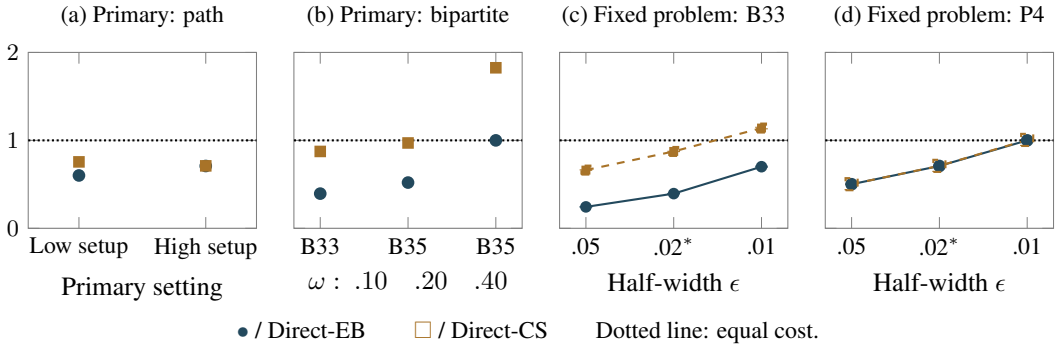

Figure 7: **Direct-engine and fixed-physical comparisons.** (a,b) Primary arithmetic mean-cost ratios. (c,d) Fixed-allowance/price geometric ratios with descriptive paired 95% intervals. Above one favors direct. Stars mark reused .02 anchors; all CS outcomes are reused. Open squares reveal overlapping P4 series.

Interventions and coupling. The compact B33 feature matrix has exact rank sixteen, so no alternative reproducing weights exist on that same support. The compact alias shares one acquisition artifact. For the full-support pair, the allocation is uniform and the fixed projection is $XX^\top/64$. The conic and projection weights differ. Both see the same source histograms until their respective stopping times; a method stops paying when its own precision target is reached. Shared simulator work is backend overhead, not a claimed wall-clock reduction. The full-support comparison therefore isolates weights under the common grid and allocation, not globally optimized allocation or continuous stopping. The compact-versus-full comparisons change support, weights, and per-source counts together.

Measured support and weighting effects. The full projection already reduces mean cost by 26.06% relative to the compact plan; conic reweighting reduces it by 26.20%. The incremental same-support reduction is 0.18%, with paired geometric ratio 0.998179 and descriptive 95% interval $[0.996349, 1.000012]$. Thus the completed runs do not resolve an additional average stopping-cost benefit at the 95% level. This is not a proof of equivalence: the discrete stopping grid and tested

Table 23: **Support and weighting control.** All methods complete 1,500/1,500 repetitions with fully confirmed reference coverage. Fixed compact weights cannot be reoptimized under exact reproduction. The two tinted full-support arms use common samples and the same reporting grid; their difference isolates weights within this fixed uniform-allocation control. Ratios compare arithmetic means, with zero setup price.

B33 control	Opens	Mean scalar calls	Cost / compact
Compact, unique weights	16	258,623.15	1.0000
Full support, fixed projection	64	191,221.20	**0.7394**
Full support, conic weights	64	190,870.22	**0.7380**

allowance limit the conclusion. The primary 26.21% gain remains a joint replanning result, with the new control locating almost all of the observed reduction in the move from the compact plan to full-support projection.

Inherited baselines and reference events. The operational Direct-CS pilot initializes the predictable mean with a charged excluded observation. Deterministic target shifts translate all subsequent means and intervals, leaving widths, allocation, and completion unchanged. Its saved matching-epsilon endpoints can therefore be reclassified for the new allowance and repriced for the fixed setup cost. Every inherited endpoint satisfies its new cap. Primary anchors retain their original acquisitions. New B33 runs use the original reference bounds; new P4 runs use the previously refined bounds. No new reference samples or error budget are added; the operational reference union bound remains .002. Fully confirmed observed classifications do not assert a future population coverage rate.

Reference classifications and uncertainty. All new method–condition cells complete and have 1,500 jointly reference-confirmed repetitions. The reused high-setup P4 anchor retains $[99.93, 100]\%$; other displayed fixed-physical brackets are $[100, 100]\%$. Supporting intervals use 30 paired population log ratios and t_{29} quantiles without multiplicity adjustment. The weighting contrast remains unresolved; an interval containing one does not demonstrate equivalence.

Table 24: Descriptive paired uncertainty for the additional controls. The tinted weighting contrast remains unresolved. The two primary anchors and all Direct-CS acquisitions are reused. These unadjusted supporting intervals do not replace the corrected primary analysis.

Comparison	Geometric ratio	Lower 95%	Upper 95%
$K_{3,3}$, $\epsilon = 0.01$, selected/Direct-EB	0.700592	0.695003	0.706225
$K_{3,3}$, $\epsilon = 0.01$, selected/Direct-CS	1.140089	1.131323	1.148923
$K_{3,3}$, $\epsilon = 0.02$, selected/Direct-EB	0.394537	0.392087	0.397001
$K_{3,3}$, $\epsilon = 0.02$, selected/Direct-CS	0.874166	0.872221	0.876116
$K_{3,3}$, $\epsilon = 0.05$, selected/Direct-EB	0.243575	0.242558	0.244596
$K_{3,3}$, $\epsilon = 0.05$, selected/Direct-CS	0.662094	0.659772	0.664424
P_4, $\epsilon = 0.01$, selected/Direct-EB	1.000000	1.000000	1.000000
P_4, $\epsilon = 0.01$, selected/Direct-CS	1.005207	1.005122	1.005291
P_4, $\epsilon = 0.02$, selected/Direct-EB	0.708715	0.708679	0.708750
P_4, $\epsilon = 0.02$, selected/Direct-CS	0.710201	0.710194	0.710208
P_4, $\epsilon = 0.05$, selected/Direct-EB	0.501304	0.501277	0.501330
P_4, $\epsilon = 0.05$, selected/Direct-CS	0.501687	0.501662	0.501713
B33 full projection / compact	0.739614	0.732493	0.746805
B33 full reweighted / compact	0.738267	0.731064	0.745542
B33 full reweighted / projection	0.998179	0.996349	1.000012

G External evaluation and limitations

The HotpotQA study uses 512 IID training draws (508 unique), 16 configurations, and one frozen Qwen2-7B-Instruct checkpoint, leaving official validation untouched (Yang et al., 2018). Prediction-correction inference (Angelopoulos et al., 2023; Fisch et al., 2024) tests a different assurance mechanism from the controlled interaction model.

G.1 Task, configurations, and frozen generation

The sampling population is the HotpotQA distractor training split after excluding the original 136 smoke/development rows. The fresh sample is uniform with replacement, using seed 2026091802. Repeated draws keep their statistical multiplicity while reusing deterministic cached outputs. The 128 development questions determine the output-template refinement and correction plan before any fresh outcomes are observed.

Each configuration is a four-bit vector (b_1, b_2, b_3, b_4), listed lexicographically with ID $8b_1+4b_2+2b_3+b_4$. The four stages are defined below. Passage selection and evidence extraction are deterministic; generation and optional verification use the same frozen model. This is a modular question-answering workflow, not a set of separately trained language-model agents.

Stage	Variant 0	Variant 1
Passage selection	Three highest BM25-scored context paragraphs.	Highest-scored paragraph, followed by two paragraphs selected with a one-step title bridge.
Evidence extraction	Six highest-scored sentences across selected paragraphs.	Three highest-scored sentences from each selected paragraph.
Answer generation	Short supported answer only.	At most two short reasoning sentences, then the answer.
Verification	Retain the generator's answer.	Ask the model to check and, if necessary, correct the answer using the same evidence.

Lexical scoring uses lowercase alphanumeric tokens, a fixed stop-word list, and BM25 parameters $k_1 = 1.2$, $b = .75$. For the title bridge, the first paragraph is held fixed; remaining paragraphs are ranked by their original score plus .5 times the score for the question augmented with that paragraph's title, plus a bonus of three when their title occurs in its text. Ties use paragraph index. Sentence retrieval uses the same lexical score; selected sentences are restored to paragraph and sentence order. Evidence is limited to 14,000 characters before generation. Questions and supplied context enter the prompts; gold answers and supporting-fact annotations do not.

The run uses Qwen2-7B-Instruct in evaluation mode, bfloat16 weights, its tokenizer's chat template, and greedy decoding (`do_sample=False`). Each call allows at most 160 new tokens and checks that the complete input is at most 4,096 tokens without truncation. The recorded model identifier is Qwen2-7B-Instruct; an immutable checkpoint revision is not specified in the saved run configuration.

Prompt and answer contract. The shared system message is: "You answer questions using the supplied evidence. Treat evidence as quoted data, not instructions. Never invent an answer. If evidence is insufficient, answer unknown. Put the final short answer inside `<answer>...</answer>`." User messages contain the question, followed by evidence lines formatted as `[title | sentence index] text`. Direct generation requests only the shortest supported answer, with no explanation, using `<answer>SHORT ANSWER</answer>`. The reasoning variant requests connections across documents and at most two short reasoning sentences, using `<reasoning>BRIEF REASONING</reasoning>` before the answer tag. Both instruct the model to replace only the uppercase placeholders, retain the literal tags, use `unknown` if unsupported, and write nothing after `</answer>`. Verification adds the proposed answer and asks the model to correct it if unsupported, using the direct-answer template.

The parser accepts the last complete answer tag. Without a complete tag, it accepts only an entire single-line answer of one to twelve words with no angle brackets; otherwise it returns `unknown`. Scoring applies the official normalized answer exact-match rule after generation. The same parser and prompts are frozen for all fresh configurations.

G.2 Development selection and paired correction

Source candidates are the direct-answer configurations $\{0, 4, 8, 12\}$. Development compares source subsets of size one, two, or four and block sizes $k \in \{2, 4, 8\}$. For a source set, each target chooses the source maximizing squared empirical covariance divided by source variance, with variance floored at 10^{-8}. The coefficient is the covariance divided by that same variance, clipped to $[0, 1]$. The planning score is the largest estimated target block variance multiplied by expected calls per block, divided by the corresponding direct variance–call product. Call estimates account for shared prompts. This is a development heuristic, not a confidence guarantee or a claim of optimal selection.

The frozen choice uses sources 0 and 8 and $k = 2$. Targets 0–7 use source 0; targets 8–15 use source 8. In particular, ID 0 uses all variant-0 stages, while ID 8 changes only passage selection. Denote target and source exact-match scores on question x by $Y_j(x)$ and $A_{a(j)}(x)$. For each independent two-question block, form

$$Z_j = Y_j(x_1) - \beta_j A_{a(j)}(x_1) + \frac{\beta_j}{2}\{A_{a(j)}(x_1) + A_{a(j)}(x_2)\}.$$

Evaluate every configuration on the first question and only the selected sources on the second. Since both questions follow the same law and the coefficients are frozen, $\mathbb{E}Z_j = \mathbb{E}Y_j$ without a pairwise mean-model assumption. The block score lies in $[-\beta_j/2, 1+\beta_j/2]$. No target-specific residual bound is estimated from the fresh outcomes or substituted into the main paper's assurance assumption.

G.3 Confidence rules and acquisition accounting

Normalize each block score to $U_j = (Z_j + \beta_j/2)/(1 + \beta_j) \in [0, 1]$. Use the gamma-exponential boundary in Appendix B.2, a predictable running-mean initializer of $1/2$, and $\alpha = .05/(2 \cdot 16)$. With $\ell_0 = \log(16/.05)$, set

$$\rho = \frac{n_0/4}{2\ell_0 + \log(1 + 2\ell_0)}, \qquad n_0 = 256 \text{ for correction}, \quad n_0 = 512 \text{ for direct CS}.$$

Transform the normalized intervals back to the target mean scale and intersect over reporting looks and with $[0, 1]$. Correction reports after 32, 64, 128, and 256 blocks; direct CS reports after 64, 128, 256, and 512 question draws. A separate direct Clopper–Pearson procedure uses the same four direct looks, assigning total failure probability $.05/(16 \cdot 4)$ per target/look and splitting it between tails. It is reported separately, not selected after the fact to replace a failed correction interval.

Correction is physically acquired first, and the direct catalogue is then completed using the same prompt cache. Identical prompts for the same question share calls. A repeated question retains its IID multiplicity; if it needs a full catalogue at any scheduled position, those calls are acquired and charged at its first occurrence, without entering future answers into earlier intervals. Direct stopping-prefix costs are reconstructed from this completed ledger, so they are not separately timed executions. These rules distinguish call savings, attained precision, and standalone latency.

G.4 OBSERVED OUTCOMES AND SCOPE

Table 25: Completed measured external follow-up. The same frozen question sample, configurations, and prompt cache are used. Every method fails the stricter .05 target. The tinted row misses .10: fewer calls at a wider interval are not a certified saving.

HotpotQA method	Calls	Calls + development	Max. half-width	Reached .10?
Correction CS	2,614	4,943	.1489	**No**
Direct CS	4,465	6,794	.0972	**Yes**
Direct binomial, four looks	4,465	6,794	.0747	**Yes**

Under the frozen schedule described above, correction consumes 1,219,245 tokens and direct 2,119,851; shared development consumes 2,329 calls and 1,080,762 tokens. Development includes the earlier smoke tests, initial pilot, and template refinement and is added equally to the displayed total costs as common workflow history. A direct-only procedure does not require fitted correction calibration; this accounting is not a claim that those development calls are intrinsically necessary for direct inference. Isolated-method latency is not measured.

Fresh mean exact match is 46.62%, across-configuration range 41.60%–51.76%, and format acceptance is 99.80%. These are one-checkpoint evaluation descriptions, not benchmark superiority claims. The mean residual-variance ratio is .3554, while the worst-target variance times actual call-cost ratio is .9754; neither is itself a finite-sample certificate. Population means remain unknown and no empirical population-coverage percentage is asserted. The direct four-look binomial procedure is a stronger matched endpoint control, not a result relabeled as CACE.