CERTIFYING REUSE FOR LLM COMPONENTS FROM LOCAL INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Under a bound on background interaction order, exact local intervention values can determine whether a specified calling rule remains optimal in an untested combination. In practice, however, each intervention value must be estimated from finite reward samples. The reconstruction can amplify these estimation errors, so testing fewer contexts need not make certification cheaper. This makes reuse certification an experimental design question: how much evidence is needed to support an untested decision, and which additional contexts reduce that cost? We study a fixed pair of functional roles whose marginal values determine a calling decision. Without a restriction on unobserved higher order effects, local interventions cannot certify an untested combination even with unlimited repetitions. Under a degree h model for the decision relevant contrasts, local measurements identify a size k target exactly, but certification has matching sample complexity of order $L_h(k)^2\Delta^{-2}\log(1/\delta)$, where $L_h(k)$ is the exact amplification of the local reconstruction. We then show that measuring at most 2^h-1 suitable proper subsets reduces the amplification to at most $2^{h+1}-1$, independent of k, without observing the target itself. A finite sample certifier turns these results into approval, a negative witness, or insufficient evidence. Controlled experiments reproduce the predicted amplification and reduce it from 97 to 7 in a quadratic eight background example.

1 INTRODUCTION

Reusable components are rarely deployed only in the contexts in which they were evaluated. Skills, tools, memories, and reasoning modules are selected alongside other resources and under background conditions that combine across tasks (Kong et al., 2026; Xia et al., 2026; Li et al., 2026). Exhaustively testing every combination quickly becomes impractical. A natural alternative is to evaluate simpler contexts and use those measurements to support decisions in more complex combinations. The central question is when such local evidence is strong enough to justify a use that was never tested directly.

Controlled behavioral tests and causal interventions reveal effects that aggregate success can leave hidden (Ribeiro et al., 2020; Geiger et al., 2021; 2025). Interaction decompositions and structured set functions show how values on simpler subsets can determine larger combinations under structural assumptions (Harsanyi, 1963; Kang et al., 2025; Butler et al., 2025), while pure exploration and experimental design study how finite samples should be allocated to answer a target decision (Fiez et al., 2019; Garivier & Kaufmann, 2016). The remaining issue in our setting is what happens when the measurements that identify an untested combination must themselves be estimated from finite samples. An exact reconstruction may combine many noisy local estimates with large alternating coefficients. Noiseless identifiability can therefore overstate how much evidence a local evaluation actually provides. The target may remain uniquely determined while its confidence interval is too wide to support a deployment decision.

To address this gap, we formulate reuse from local interventions as a finite sample certification problem. We fix two task grounded functional roles and a calling rule, then intervene on whether neither role, either role alone, or both roles are available under a background U. Rather than recover the full task value function, we track the marginal contrasts that determine the specified decision. Without restrictions on unobserved higher order effects, local interventions cannot certify an untested combination even with infinite repeated samples. Under a degree h model for the decision relevant

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.

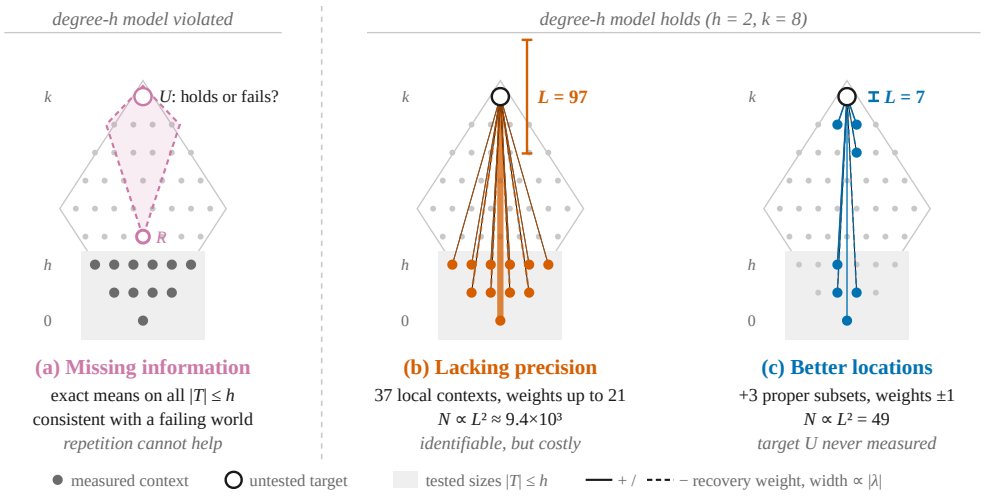

Figure 1: Preview of the central certification problem. **(a)** Without structural control, a higher order effect can leave every local intervention unchanged while changing the decision at the untested target, so repetition cannot recover the missing information. **(b)** Under a degree-h model, local measurements identify the target, but the reconstruction can strongly amplify sampling noise; for $h = 2$ and $k = 8$, $L = 97$. **(c)** Three proper subset measurements reduce this amplification to $L = 7$ without observing the target. Thus exact local identifiability need not provide a usable certificate, while a small change in measurement locations can.

contrasts, all backgrounds of size at most h identify an arbitrary target. For a target containing $k > h$ background elements, we derive the exact amplification $L_h(k)$ of this reconstruction and prove matching single target sample complexity of order $L_h(k)^2\Delta^{-2}\log(1/\delta)$. The analysis also shows how to change the experiment. Measuring a small number of larger proper subsets yields a reconstruction whose amplification no longer grows with k, without measuring the final target. We then combine these recovery formulas with valid finite sample intervals so that the evaluator can approve the calling rule, return a background that violates it, or report that the available evidence is insufficient. Figure 1 previews this progression: local evidence can fail because the target is not identified or because its reconstruction is statistically unstable, while a small change in the measured contexts can make the same unobserved target certifiable.

Our main contributions are summarized below.

- We characterize the limits and statistical cost of local reuse evidence. Without structure, unobserved higher order effects make an untested combination unidentifiable from local interventions. Under a degree h model, we derive the exact local amplification $L_h(k)$ and prove matching single target sample complexity of order $L_h(k)^2\Delta^{-2}\log(1/\delta)$.
- We show how changing the tested contexts can remove the growth with target size. Measuring at most $2^h - 1$ suitable proper subsets gives an explicit reconstruction with amplification at most $2^{h+1} - 1$, independent of k, while the final target remains unobserved.
- We turn these results into a finite sample certification procedure that approves a specified reuse decision, returns a negative witness, or reports insufficient evidence. Controlled experiments reproduce the predicted statistical cost, the benefit of targeted calibration, and the failure caused by hidden higher order effects.

2 Related Work

Work on reusable reasoning studies how components can be identified, retrieved, and recombined across tasks (Kong et al., 2026; Xia et al., 2026). SkillsBench measures how externally supplied skills affect performance across diverse tasks (Li et al., 2026). Our question begins after the reusable

components and their roles have been specified. Once a component is to be used in a background that was not directly evaluated, the remaining problem is what intervention evidence justifies that use.

Behavioral testing changes controlled aspects of an input to expose failures hidden by aggregate performance (Ribeiro et al., 2020), while causal abstraction uses interventions to relate high level causal variables to neural systems (Geiger et al., 2021; 2025). These approaches establish why interventions can reveal functional dependencies that ordinary outputs do not identify. We ask what follows when interventions cover only a subset of a combinatorial background space. The issue is then not whether an intervention is informative at a measured context, but whether finite intervention evidence supports a decision in an unmeasured one.

Möbius and Harsanyi decompositions express how set valued inputs contribute through interactions (Harsanyi, 1963), and recent work such as SPEX and ProxySPEX develops scalable recovery of sparse interactions in language models (Kang et al., 2025; Butler et al., 2025). A specified calling decision, however, depends on particular marginal contrasts rather than the complete task value function. Common background effects can cancel from these contrasts even when the full value function has higher order dependence. We therefore study when the decision relevant contrasts can be certified without recovering every interaction, and how the geometry of local reconstruction amplifies sampling uncertainty.

Pure exploration and transductive experimental design allocate samples to answer a target question without estimating every quantity equally well (Fiez et al., 2019; Garivier & Kaufmann, 2016). These results provide tools for sample allocation, stopping, and information lower bounds. In our setting, the backgrounds that may be measured determine the reconstruction geometry before samples are allocated. We characterize the statistical cost induced by a local background design and show that measuring a small set of proper subsets can change this cost without observing the target itself. Sample allocation and context selection therefore address different parts of the certification problem.

3 Interventions and the Reuse Decision

We first formalize the decision that local evidence must support. Role interventions determine the marginal value of each calling choice, while a fixed calling cost determines which choices are optimal. Certification asks whether these choices remain justified in backgrounds that have not been evaluated directly. Let $B = \{1, \dots, d\}$ denote background elements and let $U \subseteq B$ be the background present in one evaluation context. We fix two ordered functional roles P and Q before observing intervention rewards. Their concrete realization may change with the task, but their task meaning and the intervention protocol are held fixed across backgrounds.

For $S \subseteq \{P, Q\}$, let $\mathcal{T}(U, S)$ denote the intervention that makes exactly the roles in S available. With reward in $[0, 1]$, define

$$v_U(S) = \mathbb{E}[r(Y) \mid Y \sim M(\mathcal{T}(U, S))]. \tag{1}$$

The four values $v_U(\varnothing)$, $v_U(P)$, $v_U(Q)$, and $v_U(PQ)$ separate the roles' standalone and conditional contributions. The three marginal values relevant to the calling decision are

$$\alpha(U) = v_U(P) - v_U(\varnothing), \qquad \beta(U) = v_U(Q) - v_U(\varnothing), \qquad \gamma(U) = v_U(PQ) - v_U(P). \tag{2}$$

Here α is the value of calling P alone, β is the value of calling Q alone, and γ is the value of calling Q when P is already available.

Let τ be a fixed calling cost or threshold. Calling a resource with marginal value d gives surplus $d - \tau$, whereas not calling it gives zero. We study the decision pattern that uses P, rejects Q when it is alone, and uses Q when P is already available. Its signed gaps are

$$g(U) = \big(\alpha(U) - \tau,\ \tau - \beta(U),\ \gamma(U) - \tau\big), \qquad \kappa_\tau(U) = \min_j g_j(U). \tag{3}$$

The sign is retained because a negative gap is evidence against using the prescribed action in that case.

The same marginal values also define an ordered relation between the two roles. Let

$$A(U) = \alpha(U) - \beta(U), \qquad I(U) = \gamma(U) - \beta(U), \tag{4}$$

with relational margin

$$\rho(U) = [\min\{A(U), I(U)\}]_+ . \tag{5}$$

Thus $\rho(U) > 0$ when P has greater standalone value than Q and Q becomes more valuable once P is already available.

Proposition 1 (Relation and deployment decision)**.** *The calling pattern $(1, 0, 1)$ is uniquely optimal in the three decision cases at U exactly when $\kappa_\tau(U) > 0$. If the cases occur with probabilities $p_j(U)$, then the regret from using this pattern is*

$$\mathcal{R}(U) = \sum_{j=1}^{3} p_j(U) \, [-g_j(U)]_+ . \tag{6}$$

Moreover,

$$\rho(U) = 2 \left[\sup_{\tau \in \mathbb{R}} \kappa_\tau(U) \right]_+ . \tag{7}$$

The relational margin allows the threshold to be chosen, whereas deployment fixes it. A positive ρ therefore identifies a range of calling costs for which the ordered relation changes optimal actions, but certification must resolve the gaps at the cost that will actually be used. The same relational margin can correspond to disjoint valid threshold intervals; Appendix B gives a concrete example. These gaps specify what must be certified, but their values are not observed in an untested background. We next ask whether interventions on smaller backgrounds determine their signs, and how much evidence is required when they do.

4 The Limits and Statistical Cost of Local Evidence

Local evidence can fail for two different reasons, illustrated in Figure 1a–b. The measured backgrounds may not determine the target at all, or they may determine it through a reconstruction that is sensitive to sampling noise. We first separate these cases and then quantify the cost of the second.

4.1 What local interventions cannot identify

Suppose the evaluator can test only backgrounds with at most h elements. Repeating these tests can remove sampling error, but it cannot reveal an effect that occurs only when a larger combination is present.

Theorem 1 (Undetected higher order failure)**.** *Suppose only backgrounds U with $|U| \leq h < d$ may be tested, with arbitrary precision and all four role conditions. Without a restriction on higher order background effects, no procedure can both approve every system on which the calling rule is valid for all backgrounds and avoid approving some system on which the rule fails at an untested background.*

The proof places the entire difference between two systems on one set of $h + 1$ background elements. Every allowed intervention therefore has the same distribution in both systems, while the prescribed calling rule is valid in one world and fails when that larger set appears in the other. The obstruction is informational rather than statistical. More repetitions cannot repair it. A positive result must instead restrict the effects that remain invisible to the local query family. Appendix A gives the explicit construction.

4.2 A structural condition on decision relevant contrasts

Let

$$\mathcal{H}_h = \{R \subseteq B : |R| \leq h\}. \tag{8}$$

We assume that each marginal value in (2) has degree at most h in the background indicators,

$$f(U) = \sum_{R \in \mathcal{H}_h} \theta_R \mathbf{1}\{R \subseteq U\}. \tag{9}$$

The assumption is on α, β, and γ, which determine the calling decision, rather than on the complete task value. In particular, the common baseline $b(U) = v_U(\varnothing)$ may contain arbitrary higher order

dependence because it cancels from these contrasts. The decision relevant structure can therefore be simple even when the full behavior is not. Values on backgrounds of size at most h determine all coefficients in (9). Substituting those coefficients into a larger target yields the following direct reconstruction.

Theorem 2 (Local interpolation and exact amplification)**.** *Let f satisfy* (9)*, and suppose $f(T)$ is observed for every T with $|T| \le h$. For a target U with $k = |U| > h$,*

$$f(U) = \sum_{\substack{T \subseteq U \\ |T| \le h}} \lambda_T(U) f(T), \qquad \lambda_T(U) = (-1)^{h-|T|} \binom{k-|T|-1}{h-|T|}. \tag{10}$$

The local recovery weights are unique, and their absolute sum is

$$L_h(k) = \sum_{t=0}^{h} \binom{k}{t} \binom{k-t-1}{h-t}. \tag{11}$$

In particular,

$$L_1(k) = 2k-1, \qquad L_2(k) = 2k^2 - 4k + 1, \tag{12}$$

and for fixed h, $L_h(k) = \Theta(k^h)$.

Theorem 2 gives exact identification when the local values are known without error. Its coefficients, however, can be large and alternate in sign. If every local estimate has absolute error at most ϵ, then the same reconstruction only guarantees

$$|\widehat{f}(U) - f(U)| \le L_h(k)\epsilon. \tag{13}$$

Exact recovery relies on cancellation between the local means, while their estimation errors need not cancel. Identifiability alone therefore does not determine whether the target can be certified from a finite observation budget.

4.3 Identification can be statistically expensive

We isolate this cost for one unmeasured scalar target. Each allowed query returns an independent Bernoulli reward from a background T with $|T| \le h$. The target is promised to lie at least Δ above or at least Δ below a fixed decision boundary. A procedure is δ reliable if it returns the correct side with probability at least $1-\delta$ for every promised instance.

Theorem 3 (Matching local sample complexity)**.** *For degree at most h functions with values in $[1/4, 3/4]$, a target U of size $k > h$, and local queries restricted to $|T| \le h$, any sequential δ reliable procedure requires, on some instance,*

$$\mathbb{E}N = \Omega\left(\frac{L_h(k)^2}{\Delta^2} \log \frac{1}{\delta}\right). \tag{14}$$

Conversely, allocating local samples in proportion to $|\lambda_T(U)|$ attains

$$N = O\left(\frac{L_h(k)^2}{\Delta^2} \log \frac{1}{\delta}\right) \tag{15}$$

up to the additive cost of integer rounding.

The upper bound assigns observations in proportion to the absolute recovery weights, which makes the estimation variance scale as $L_h(k)^2/N$. For the lower bound, two degree h systems are constructed so that every local Bernoulli mean differs by only order $\Delta/L_h(k)$ while their target values lie on opposite sides of the decision boundary. Even an adaptive local procedure must accumulate enough information to distinguish these worlds. Appendix A gives both arguments. The lower bound already allows adaptive choice among local contexts. The $L_h(k)^2$ factor is therefore a property of the local measurement geometry, not an artifact of uniform sampling. For $h = 2$ and $k = 8$, all $1+8+\binom{8}{2} = 37$ local backgrounds identify the target exactly, but $L_2(8) = 97$, giving squared amplification 9409. This motivates changing the measurement locations rather than only redistributing samples across the same local design.

5 STABLE CALIBRATION WITHOUT MEASURING THE TARGET

The preceding lower bound holds while every query remains local. Figure 1c suggests the alternative: measure a few larger proper subsets of the target without observing the target itself. We ask whether these additional locations can reduce uncertainty. The quadratic case makes the mechanism explicit. For a degree at most two function and distinct $a, b \in U$, the mixed difference in a and b is unchanged by the remaining background. Equating this difference at the empty background and at $U \setminus \{a, b\}$ gives

$$\begin{aligned} f(U) = f(U \setminus \{a\}) + f(U \setminus \{b\}) - f(U \setminus \{a, b\}) \\ + f(\{a, b\}) - f(\{a\}) - f(\{b\}) + f(\varnothing). \end{aligned} \tag{16}$$

Every coefficient has magnitude one. At $k = 8$, the local design has amplification 97, while this seven context representation has amplification 7. If the complete local design has already been measured, only the three larger backgrounds $U \setminus \{a\}$, $U \setminus \{b\}$, and $U \setminus \{a, b\}$ must be added. If only this target matters, the seven contexts in (16) can instead be measured directly. The same finite difference argument extends to any declared degree.

Theorem 4 (Complement calibration)**.** *Let f have background degree at most $h \geq 1$, let $|U| > h$, and choose an anchor set $J \subseteq U$ with $|J| = h$. Then*

$$f(U) = \sum_{\varnothing \neq T \subseteq J} (-1)^{|T|+1} f(U \setminus T) + \sum_{V \subseteq J} (-1)^{h-|V|} f(V). \tag{17}$$

The target U is never measured. Relative to the complete local design, at most $2^h - 1$ additional proper subset backgrounds are needed. After duplicate contexts are merged, the absolute weight sum is at most $2^{h+1} - 1$, independent of $|U|$.

The proof compares the same h fold finite difference at two backgrounds. Under degree at most h, only the coefficient involving all anchors survives, so the two differences agree. Isolating the target term gives (17). Appendix A gives the full derivation.

Corollary 1 (Target certification from proper subsets)**.** *Under independent bounded rewards and a target separated from the decision boundary by Δ, the complement design certifies one degree at most h target using*

$$O\left(\frac{4^h}{\Delta^2} \log \frac{1}{\delta}\right) \tag{18}$$

reward observations, up to constants determined by the intervention means in the target contrast. The bound is independent of the target size k.

The complement design removes the growth with k, but the interaction order h, the decision gap Δ, and the confidence level still control the cost. It also targets only the requested decision rather than recovering the complete degree h function, which has $\sum_{j=0}^{h} \binom{k}{j}$ free coefficients. Appendix C compares the explicit construction with richer proper subset designs. The recovery formulas now specify which intervention means are needed and how their uncertainty combines. We next turn these representations into finite sample decisions over a deployment family.

6 FINITE SAMPLE REUSE CERTIFICATES

For a fixed finite deployment family, the evaluator should approve only when every required decision gap is positive, return a background when one gap is negative, and otherwise retain an unresolved conclusion. We first construct confidence intervals for these gaps and then allow the sampling allocation to adapt across phases.

6.1 FROM MEASUREMENTS TO CERTIFIED DECISIONS

Index allowed scalar reward measurements by q, where one measurement is one background together with one of the four role conditions. Index the deployment gaps to be certified by z. Every such gap can be written as

$$g_z = b_z + \sum_q c_{zq} \mu_q, \tag{19}$$

where the coefficients follow from a local or calibrated recovery representation.

At phase t, the certifier chooses sample counts n_{tq} from the history of earlier phases, freezes this allocation, and then draws fresh independent rewards. For Z deployment gaps, define

$$\widehat{g}_{tz} = b_z + \sum_q c_{zq}\widehat{\mu}_{tq},$$
$$r_{tz} = \sqrt{\frac{1}{2}\log\frac{2Zt(t+1)}{\delta}\sum_q \frac{c_{zq}^2}{n_{tq}}}. \tag{20}$$

Intervals are intersected across phases. The procedure approves when every lower endpoint is positive, refutes when one upper endpoint is negative and returns that background, and otherwise continues until the budget is exhausted, in which case it reports insufficient evidence.

Theorem 5 (Sound adaptive certification)**.** *Under the exact declared contrast model and independent rewards, with probability at least $1-\delta$, every interval constructed from* (20) *at every phase contains its true deployment gap. Consequently, every approval is valid simultaneously over the specified deployment family and every returned negative witness is genuine. If the lower endpoints for the three cases at U are $\ell_j(U)$, then*

$$\mathcal{R}(U) \le \sum_j p_j(U)\left[-\ell_j(U)\right]_+. \tag{21}$$

With geometrically growing budgets and a positive allocation floor, the procedure terminates on the coverage event whenever all relevant gaps are separated from zero by a fixed positive margin, or at least one gap is separated below zero by such a margin.

Conditioned on the history before a phase, the new allocation is fixed and the fresh observations satisfy the weighted concentration bound. A summable error budget over phases and deployment gaps gives simultaneous coverage. The guarantee therefore permits any allocation chosen from earlier phases. The active allocation concentrates observations on unresolved measurements that influence near boundary gaps while retaining a small uniform floor. Appendix D gives the allocation rule and protocol.

6.2 Allowing bounded structural error

The preceding certificate is conditional on the declared contrast model. If $f = f_0 + e$, with f_0 in the declared class and $|e| \le \eta$ on measured and target contexts, a recovery vector a obeys

$$|f(U_*) - a^\top \widehat{f}| \le \sum_i |a_i|\epsilon_i + (1 + \|a\|_1)\eta, \tag{22}$$

where ϵ_i bounds measurement error. The same amplification that controls statistical noise therefore also controls worst case structural error. A residual bound supplies information beyond the local observations. When higher order effects are unrestricted, Theorem 1 shows that those observations cannot establish such a bound by themselves.

7 Experiments

The experiments test three consequences of the theory. We ask whether changing measurement locations improves certification at a fixed budget, whether adaptive allocation reduces the cost of simultaneous decisions when the query family is fixed, and whether hidden higher order effects invalidate conclusions drawn from a misspecified local model. We use a synthetic environment in which role P performs hexadecimal decoding and role Q performs aggregation. Task subtypes may require either role, both roles, or neither, while background elements trigger handler conflicts or repairs. A separate role independent branch adds higher order variation to the common baseline, so the complete task value can be more complex than the decision relevant contrasts. The declared contrast degree is controlled by the generator rather than inferred by the certifier.

Task outcomes are enumerated to obtain exact success probabilities. Repeated noisy measurements are then sampled through the corresponding binomial sufficient statistics. Reported budgets count simulated independent reward observations. Appendix D gives the generator, exact reconstruction checks, and complete parameter grids.

Table 1: Eight background quadratic target. The approval fraction is the fraction of 1,000 repetitions whose confidence interval is sufficient to certify the true positive target. Budgets differ only by integer allocation rounding.

Design	RMSE at 10^6	Approval at 10^5	Approval at 10^6
Uniform local	.08539	.005	.015
Optimal local	.04742	.006	.040
Complement	.00356	.954	1.000
Direct target	.00049	1.000	1.000

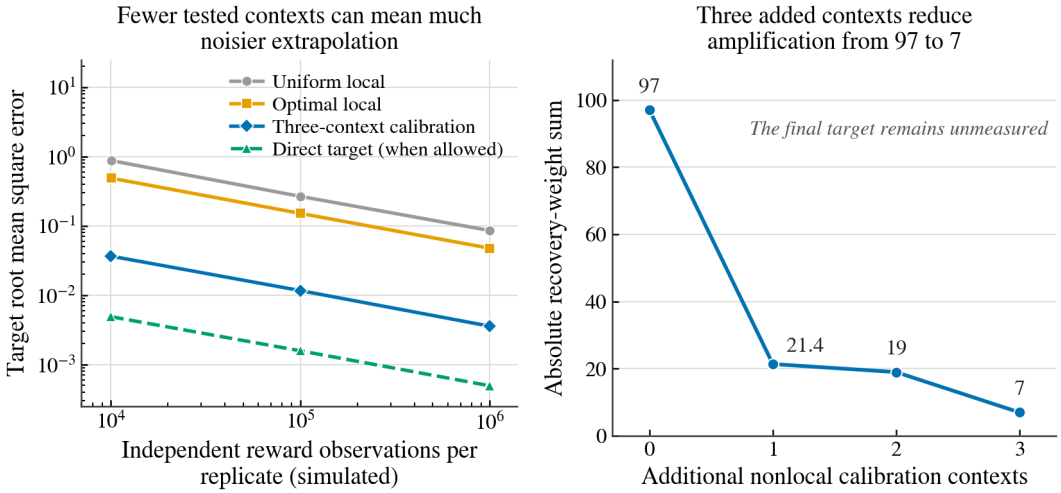

Figure 2: Left: target RMSE as the simulated reward budget grows for measurement designs that recover the same target. Right: starting from all 37 quadratic local backgrounds, three added proper subsets reduce amplification from 97 to 21.4, 19, and 7 while the final target remains unobserved. The left panel measures the seven context complement formula from scratch, whereas the right panel begins from the complete local design and adds larger proper subsets.

7.1 Measurement design changes certification cost

We first fix one unmeasured quadratic target with eight background elements and compare the evidence obtained from four measurement designs over 1,000 repetitions per setting. The local designs measure all backgrounds of size at most two. The optimally allocated local design uses the fixed target allocation from Theorem 3. The complement design uses the seven contexts in (16) and never measures the final target. Direct target measurement is included as a stronger information reference.

The empirical errors follow the predicted amplification. At roughly 10^6 observations, predicted and observed RMSE are .04850 and .04742 for optimal local sampling and .00350 and .00356 for complement calibration. This reduction changes the certification outcome at the same observation budget. At roughly 10^5 observations, the complement design certifies the target in 95.4% of repetitions, while the optimally allocated local design does so in $.6\%$.

These comparisons change which backgrounds can be measured. We next hold the local query family fixed and vary how observations are allocated within it.

7.2 Sample allocation for simultaneous certification

The second experiment uses degree one contrasts with ten background elements and simultaneously certifies decisions over all $1{,}024$ deployment backgrounds. The local methods can measure only the empty background and ten singleton backgrounds. We compare uniform local sampling, the phased active allocation from Section 6, and a uniform exhaustive reference that directly measures

Table 2: Median simulated reward observations, in millions, for simultaneous certification over $1{,}024$ backgrounds. Every definite decision in the 600 safe and unsafe repetitions is correct. All 300 near boundary repetitions remain unresolved at the budget cap.

Regime	Uniform local	Active local	Uniform direct
Safe, minimum gap .07	41.142	14.234	69.988
Unsafe, minimum gap $-.10$	24.200	4.924	24.240
Near boundary, minimum gap .01	80.000	80.000	79.999

all backgrounds. Each method is evaluated over 100 independent repetitions in three regimes with confidence parameter $.05$ and an 80 million observation cap.

The active allocation reduces median sample use relative to uniform local sampling by about 2.89 times in the safe regime and 4.92 times in the unsafe regime. All methods remain unresolved at the observation cap when the minimum gap is $.01$, because the corresponding confidence intervals still include zero. The exhaustive reference uses a uniform rather than sequentially optimized allocation.

7.3 Hidden interactions limit local certification

The preceding experiments satisfy the declared contrast model. We finally keep every quadratic local intervention value unchanged while adding a third order effect that appears only in larger combinations. Across 40 such systems, every quadratic local intervention exactly matches a safe system, yet 32 of the 256 backgrounds violate the calling rule. A certifier that incorrectly assumes quadratic contrasts approves all 40 systems. The failure persists even with exact local means, so additional repetitions at the same backgrounds cannot resolve it. Appendix D gives the construction and the corresponding failure visualization.

8 Discussion and Conclusion

The central implication is that local identifiability is not itself a finite sample certificate. When unobserved higher order effects remain unrestricted, the tested contexts do not contain enough information and repeating the same interventions cannot resolve the missing combination. When a structural model does identify the target, the remaining problem is statistical. An unstable reconstruction can require many observations even though the noiseless target is uniquely determined. These failures call for different responses: missing information requires new structural evidence or new measurement locations, while sampling uncertainty can be reduced through additional observations.

For reuse evaluation, this distinction shifts attention from reconstructing complete behavior to supporting a specified decision. The relational margin describes whether some threshold can make the two role distinctions jointly relevant, while deployment fixes a particular threshold and asks whether its signed gaps remain positive. Context selection and sample allocation then affect different parts of the evidence problem. Better allocation reduces uncertainty within a chosen measurement design, whereas proper subset calibration changes the geometry of the reconstruction itself. Direct target measurement remains the strongest option when it is available and affordable.

Within the declared decision contrast class, we characterize the information obstruction, the matching local sample complexity for a single target, and a proper subset design whose amplification no longer grows with target size. Under independent rewards and a finite deployment family, the same recovery representations yield valid approval, refutation, or unresolved conclusions. The controlled experiments verify these statistical mechanisms and show the failure caused by hidden interactions when the structural condition is wrong. The present results do not establish that language model behavior generally satisfies the degree condition. They instead give a criterion for deciding whether limited intervention evidence is missing information, lacking precision, or measured in the wrong locations.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript, to check the correctness and internal consistency of mathematical derivations and proofs, and to assist with experimental plotting.

REPRODUCIBILITY STATEMENT

The theoretical assumptions, recovery formulas, and finite sample guarantees are specified in Sections 3–6. The synthetic environment, degree bounds, background sizes, measurement budgets, repetition counts, confidence parameters, allocation rules, and parameter grids are specified in Section 7 and Appendix D.

REFERENCES

Landon Butler, Abhineet Agarwal, Justin Singh Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. ProxySPEX: Inference-efficient interpretability via sparse feature interactions in LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2426. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/68dced1638d71d4d21598f79ae91262e-Abstract-Conference.html.

Tanner Fiez, Lalit Jain, Kevin Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, 2016.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963. doi: 10.2307/2525487.

Justin Singh Kang, Landon Butler, Abhineet Agarwal, Yigit Efe Erginbas, Ramtin Pedarsani, Bin Yu, and Kannan Ramchandran. SPEX: Scaling feature interaction explanations for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28878–28903, 2025. URL https://proceedings.mlr.press/v267/kang25a.html.

Lingjing Kong, Xin Liu, Guangyi Chen, Martin Q. Ma, Xiangchen Song, Yuekai Sun, Mikhail Yurochkin, Taylor W. Killian, Ruslan Salakhutdinov, Kun Zhang, Eric P. Xing, and Zhengzhong Liu. From reasoning traces to reusable modules: Understanding compositional generalization in language model reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Xiangyi Li, Yimin Liu, Wenbo Chen, et al. SkillsBench: Benchmarking how well agent skills work across diverse tasks. *arXiv preprint arXiv:2602.12670*, 2026.

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

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, Zeyu Zheng, Cihang Xie, and Huaxiu Yao. SkillRL: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

A Proofs of the Main Results

A.1 Proof of Proposition 1

In a case with marginal value d, activating the resource has surplus $d - \tau$ and not activating has surplus zero. The prescribed actions $(1, 0, 1)$ are therefore uniquely optimal exactly when

$$\alpha - \tau > 0, \qquad \tau - \beta > 0, \qquad \gamma - \tau > 0, \tag{23}$$

which is equivalent to $\kappa_\tau > 0$. Choosing the wrong action in one case loses $|d - \tau|$. Weighting the losses by the case probabilities gives (6).

For the relation to hold, $A = \alpha - \beta > 0$ and $I = \gamma - \beta > 0$. The largest decision margin obtainable by choosing τ is

$$\begin{aligned}\sup_\tau \min\{\alpha - \tau, \tau - \beta, \gamma - \tau\} &= \sup_\tau \min\{\min(\alpha, \gamma) - \tau, \tau - \beta\} \\ &= \frac{\min(\alpha, \gamma) - \beta}{2} = \frac{\min(A, I)}{2}.\end{aligned} \tag{24}$$

Taking the positive part yields (7).

A.2 Proof of Theorem 1

Choose a set $R \subseteq B$ with $|R| = h + 1$. Consider a mixture of deterministic tasks with masses .10 solvable without either role, .30 requiring only P, .05 requiring only Q, .35 requiring both, and .20 unsolvable. In the first system the joint tasks succeed whenever both roles are available. In the second they fail exactly when $R \subseteq U$. The corresponding value vectors are

$$(.10, .40, .15, .80) \tag{25}$$

for the first system and

$$(.10, .40, .15, .80 - .35\mathbf{1}\{R \subseteq U\}) \tag{26}$$

for the second. Every allowed test background has size at most h, so the vectors coincide under all allowed interventions. Any adaptive transcript therefore has the same law in both systems.

At $\tau = .20$, the first system has gaps $(.10, .15, .20)$ everywhere. At $U = R$, the second has gaps $(.10, .15, -.15)$. An approval rule must consequently have the same approval probability in a safe and an unsafe world, so it cannot simultaneously approve every safe system with high probability and avoid approval on every unsafe one.

A.3 A general linear identification lemma

The following lemma is useful for understanding why the recovery coefficients in the main text are the relevant objects.

Lemma 1 (Exact linear identification)**.** *Let measured scalar contrasts be $y = X\theta$ and let the target be $f_* = x_*^\top \theta$. Over the unrestricted real linear parameter class, f_* is identifiable from the noiseless measurements exactly when x_* belongs to the row span of X. In that event every a satisfying $X^\top a = x_*$ gives $f_* = a^\top y$. Otherwise two parameters give the same measurements and different targets.*

Proof. If $X^\top a = x_*$, then $a^\top X\theta = x_*^\top \theta$. Conversely, if x_* is outside the row span, there is a vector z in the null space of X with $x_*^\top z \neq 0$. The parameters θ and $\theta + tz$ produce the same measurements and different targets. For a bounded reward model containing an interior feasible parameter, t can be chosen small enough to preserve feasibility. □

A.4 Proof of Theorem 2

Möbius inversion on the downward closed local family gives

$$\theta_R = \sum_{T \subseteq R} (-1)^{|R|-|T|} f(T), \qquad |R| \le h. \tag{27}$$

For a target U,

$$f(U) = \sum_{\substack{R \subseteq U \\ |R| \le h}} \theta_R = \sum_{\substack{T \subseteq U \\ |T| \le h}} f(T) \sum_{r=|T|}^{h} (-1)^{r-|T|} \binom{k-|T|}{r-|T|}. \tag{28}$$

The alternating partial binomial identity

$$\sum_{j=0}^{m} (-1)^j \binom{n}{j} = (-1)^m \binom{n-1}{m} \tag{29}$$

for $m < n$ gives (10). The local measurement matrix is square under the subset ordering and triangular with ones on the diagonal, so the recovery weights are unique.

Taking absolute values and grouping terms by $t = |T|$ gives (11). Direct calculation yields $L_1(k) = 2k - 1$ and $L_2(k) = 2k^2 - 4k + 1$. For fixed h, the term of order k^h obtained by expanding the finite sum has leading coefficient $2^h/h!$, which gives the stated asymptotic order.

A.5 Proof of Theorem 3

For the upper bound, let $\widehat{f}(T)$ be the empirical mean from n_T independent Bernoulli samples. Weighted Hoeffding gives

$$\mathbb{P}\left(\left| \sum_T \lambda_T (\widehat{f}(T) - f(T)) \right| \ge t \right) \le 2 \exp\left(- \frac{2t^2}{\sum_T \lambda_T^2 / n_T} \right). \tag{30}$$

Choose

$$n_T = \left\lceil n \frac{|\lambda_T|}{L} \right\rceil. \tag{31}$$

Then $\sum_T \lambda_T^2 / n_T \le L^2/n$. An incorrect sign decision under the promised gap requires an error of at least Δ. Taking

$$n \ge \frac{L^2}{2\Delta^2} \log \frac{2}{\delta} \tag{32}$$

gives the upper bound, with at most one extra sample per measured context from integer rounding.

For the lower bound, write $s_T = \operatorname{sign}(\lambda_T)$ and define local means

$$f_\pm(T) = \frac{1}{2} \pm \frac{\Delta}{L} s_T. \tag{33}$$

They extend uniquely to degree h functions, and interpolation gives

$$f_\pm(U) = \frac{1}{2} \pm \Delta. \tag{34}$$

The construction remains in $[1/4, 3/4]$ for $\Delta \le 1/4$. Every allowed local query sees Bernoulli means $1/2 \pm a$ in one order, where $a = \Delta/L$. Its KL divergence is at most $16a^2$. The chain rule for an adaptive stopped experiment therefore gives

$$\mathrm{KL}(\mathbb{P}_+, \mathbb{P}_-) \le 16 \frac{\Delta^2}{L^2} \mathbb{E}_+ N. \tag{35}$$

Data processing through the final binary decision requires transcript KL at least $\mathrm{kl}(1-\delta, \delta)$. Hence

$$\mathbb{E}_+ N \ge \frac{L^2}{16\Delta^2} \mathrm{kl}(1-\delta, \delta), \tag{36}$$

which proves the lower order. The same argument applies with the two worlds exchanged.

A.6 Proof of Theorem 4

For $J \subseteq U$ with $|J| = h$, the h fold discrete derivative $\Delta_J f(U \setminus J)$ contains only coefficients of monomials that include all elements of J. Under degree at most h, the only such coefficient is θ_J. The same derivative at the empty background equals the same coefficient. Therefore

$$\sum_{T \subseteq J} (-1)^{|T|} f(U \setminus T) = \sum_{V \subseteq J} (-1)^{h-|V|} f(V). \tag{37}$$

Isolating the term $T = \varnothing$ gives (17). The second sum uses only local contexts, while the first adds at most $2^h - 1$ proper subsets of U. Before duplicate contexts are merged, the formula has at most $2^{h+1} - 1$ coefficients of magnitude one, so its absolute weight sum is at most that value.

A.7 Proof of Corollary 1

Allocate samples in proportion to the absolute recovery coefficients in (17). Applying (30) with total absolute weight at most $2^{h+1} - 1$ gives sample order

$$O\left(\frac{(2^{h+1}-1)^2}{\Delta^2} \log \frac{1}{\delta}\right) = O\left(\frac{4^h}{\Delta^2} \log \frac{1}{\delta}\right). \tag{38}$$

If a marginal contrast requires two independently measured intervention conditions, the coefficient sum and constant adjust accordingly, but the dependence on h, Δ, and δ remains the same.

A.8 Proof of Theorem 5

Condition on the history before phase t. The allocation n_{tq} is fixed at that point. Weighted Hoeffding gives failure probability at most

$$\frac{\delta}{Zt(t+1)} \tag{39}$$

for one deployment gap when the radius is chosen as in (20). A union bound over the Z gaps and all phases uses

$$\sum_{t \geq 1} \frac{1}{t(t+1)} = 1, \tag{40}$$

so every interval contains its true gap with probability at least $1 - \delta$. Intersecting intervals across phases preserves coverage on this event. Approval and refutation are therefore sound.

For the regret statement, if $g_j(U) \geq \ell_j(U)$, then $[-g_j(U)]_+ \leq [-\ell_j(U)]_+$. Apply Proposition 1 and sum over cases.

For termination, suppose a forced allocation floor gives $n_{tq} \geq \zeta B_t / Q$ and the phase budgets B_t grow geometrically. For every fixed contrast,

$$r_{tz}^2 \leq \frac{Q\|c_z\|_2^2}{2\zeta B_t} \log \frac{2Zt(t+1)}{\delta} \longrightarrow 0. \tag{41}$$

If all true gaps are at least $\Delta > 0$, a sufficiently late phase has lower endpoints above zero. If one true gap is at most $-\Delta$, a sufficiently late phase has its upper endpoint below zero. A finite budget or a zero gap may legitimately remain unresolved.

B Relational Structure, Decision Value, and Persistence

This section records complementary properties of the same intervention values. They clarify the relation between a fixed deployment decision and the broader relational quantity, why ordinary outputs do not identify that relation, and what can be guaranteed when background damage bounds are available.

B.1 A FIXED THRESHOLD IS STRICTER THAN RELATIONAL EXISTENCE

The value vectors $(.4, .6, .4, .8)$ and $(0, .4, .2, .8)$ both have $A = I = \rho = .2$ and the same joint value .8. Their valid threshold intervals for the $(1, 0, 1)$ action pattern are $(0, .2)$ and $(.2, .4)$, which do not overlap. Thus a relation can persist while a fixed calling rule becomes invalid. Equation (7) captures the distinction by optimizing the deployment margin over the threshold before taking its positive part.

B.2 ORDINARY OUTPUTS NEED NOT IDENTIFY THE RELATION

Complete knowledge of ordinary output behavior does not in general identify the four counterfactual values in (1). For example, two systems can be assigned the same ordinary success distribution with success probability .8 while having intervention values

System	$v(\varnothing)$	$v(P)$	$v(Q)$	$v(PQ)$	A	I	ρ
M_1	.10	.40	.20	.80	.20	.30	.20
M_2	.10	.45	.45	.80	.00	.00	.00

Statistics computed only from the ordinary output distribution can coincide even though the relation differs. This is the basic reason interventions are required before the local versus untested composition question arises.

B.3 VALUE OF PRESERVING RELATIONAL DISTINCTIONS

Let a decision state T have marginal value $D = d_T(C)$ and let an abstraction $Z = q_C(T)$ merge some states. For threshold distribution μ_c, define

$$G_{\mu_c}(d) = \mathbb{E}_{\tau\sim\mu_c}[(d - \tau)_+]. \tag{42}$$

The Bayes regret of the abstraction is

$$R_\mu(q) = \mathbb{E}[G_{\mu_C}(D)] - \mathbb{E}[G_{\mu_C}(\mathbb{E}[D \mid C, Z])], \tag{43}$$

which is a Jensen gap because G_{μ_c} is convex.

For two merged cases a, b with probabilities p_a, p_b, values d_a, d_b, and

$$\bar{d}_{ab} = \frac{p_a d_a + p_b d_b}{p_a + p_b}, \tag{44}$$

the exact local regret is

$$R_{a\sim b} = p_a G(d_a) + p_b G(d_b) - (p_a + p_b) G(\bar{d}_{ab}). \tag{45}$$

If thresholds are uniform on $[\ell, u]$ with width W, and the values lie in this interval,

$$R_{a\sim b} = \frac{1}{2W} \frac{p_a p_b}{p_a + p_b} (d_a - d_b)^2. \tag{46}$$

Merging the standalone P and Q cases therefore costs a term proportional to A^2, while merging Q alone with Q after P costs a term proportional to I^2. This result studies the value of information across encountered decision states. It is distinct from the deployment regret in (6), which evaluates a fixed calling pattern in a given background.

B.4 A SUFFICIENT PERSISTENCE BOUND WHEN BACKGROUND DAMAGE IS KNOWN

The original persistence analysis considered a common set function F with

$$v_U(S) = F(U \cup S) \tag{47}$$

for background U and fixed role positions. If adding background element k can reduce asymmetry and complementarity by at most a_k and b_k at every intermediate background, then for any U

$$A(U) \geq A(\varnothing) - \sum_{k\in U} a_k, \tag{48}$$

$$I(U) \geq I(\varnothing) - \sum_{k\in U} b_k, \tag{49}$$

and hence

$$\rho(U) \geq \left[\min \left\{ A_0 - \sum_{k \in U} a_k, I_0 - \sum_{k \in U} b_k \right\} \right]_+ . \tag{50}$$

The proof adds background elements one at a time and telescopes the decrement bounds. This guarantee assumes the elementwise damage bounds are already available. The main text addresses a different question. It asks what can be inferred about an unmeasured combination from a restricted set of noisy interventions.

Möbius coefficients make the connection explicit. If $m(S)$ denotes the coefficient of F, then

$$\delta_I(k; U) = \sum_{R \subseteq U} m(R \cup \{i, j, k\}), \tag{51}$$

$$\delta_A(k; U) = \sum_{R \subseteq U} \left[m(R \cup \{i, k\}) - m(R \cup \{j, k\}) \right]. \tag{52}$$

Thus higher order background interactions are exactly the terms that can move the relational contrasts. The impossibility result in the main text can be read as the finite measurement counterpart. If such interactions are unrestricted and never activated by the tested backgrounds, local data cannot rule them out.

C Additional Certification Details

C.1 Sharp linear confidence intervals

Let a declared linear model have parameter θ, observed values y, measurement matrix X, and coordinatewise confidence radii r. Define

$$\mathcal{P} = \{\theta : |X\theta - y| \leq r\}. \tag{53}$$

For an identified target $x_*^\top \theta$, its sharp lower bound over this confidence set is

$$\inf_{\theta \in \mathcal{P}} x_*^\top \theta = \max_{a : X^\top a = x_*} \left(a^\top y - \sum_i r_i |a_i| \right). \tag{54}$$

The upper bound is the corresponding minimum of $a^\top y + \sum_i r_i |a_i|$. This follows from finite dimensional linear programming duality. It gives a direct method for finding the most stable recovery vector when multiple valid representations are available.

C.2 Alternative proper subset designs

The explicit complement identity is designed to use only a small number of additional contexts. It is not the minimum amplification over every possible proper subset design. In the eight background quadratic example, allowing all 255 proper subsets in the linear program above gives absolute weight sum $7/3$, while direct measurement of the target has weight one. These comparisons separate the guarantee of the explicit construction from the optimum under richer query permissions.

C.3 Structural misspecification

Suppose $f = f_0 + e$, where f_0 belongs to the declared feature class and $|e(U)| \leq \eta$ on measured and target contexts. If a exactly recovers $f_0(U_*)$ and the measured values obey $|y_i - f(U_i)| \leq \epsilon_i$, then

$$|f(U_*) - a^\top y| \leq \sum_i |a_i| \epsilon_i + (1 + \|a\|_1)\eta. \tag{55}$$

This is obtained by adding and subtracting the exact recovery of f_0 and applying the triangle inequality. If a separately justified anchored residual model has zero structural error at all queried contexts and at most η at the target, the structural term reduces to η.

The bound clarifies why stable measurement design matters even when the declared model is only approximate. A large absolute coefficient sum amplifies both sampling error and worst case structural residuals.

D Experimental Details

D.1 Synthetic environment

The synthetic generator contains task types whose successful solution requires no role, only P, only Q, or both. The roles are functional transformations rather than labels attached to a value table. Background elements control whether particular handlers collide, are repaired, or leave the role behavior unchanged. A separate nuisance branch changes all four intervention values together and can contain higher order background dependence. This construction allows the complete value function to violate the declared degree while preserving low degree marginal contrasts.

All task outcomes are enumerated to compute exact intervention means. Noisy evaluations use binomial sufficient statistics from these means, retaining exact ground truth for every background while permitting repeated sampling experiments.

As a construction check, we generated 40 independent quadratic environments with eight background elements. The 37 local backgrounds reconstruct the three deployment gaps on all 219 unmeasured backgrounds with maximum absolute error 8.78×10^{-15}. Applying the same quadratic reconstruction to the role independent baseline gives maximum error $.24$. This verifies that the generator can preserve low degree decision contrasts while allowing more complex complete task values.

D.2 Noise experiment design

The noise study covers background sizes $3, 4, 6, 8$, degree bounds one and two, budgets near 10^4, 10^5, and 10^6, and four measurement designs. Every configuration is repeated 1,000 times. The confidence intervals use the same independent reward model as the theory. Direct target measurement is included only as a reference for the information available when the final combination itself may be tested.

For the key eight background quadratic setting, the theoretical and empirical RMSE at a budget near 10^6 are

Design	Predicted	Empirical
Uniform local	.083678	.085386
Optimal local	.048499	.047423
Complement	.003500	.003558
Direct target	.000500	.000491

The empirical 95% interval coverage in these four conditions is $.995$, $.998$, $.992$, and $.991$, respectively.

D.3 Greedy calibration path

Starting from the 37 local backgrounds in the eight background quadratic problem, we use the linear program in (54) to evaluate candidate additional proper subsets. A greedy path adds three contexts and changes the absolute coefficient sum from 97 to 21.4, then 19, then 7. The final target is not among the measured contexts. This experiment illustrates how added query permissions change recovery geometry. The explicit complement identity in Theorem 4 obtains the same bound of seven without relying on this greedy search.

D.4 Simultaneous certification protocol

The multi background study uses degree one contrasts with ten background elements. The deployment family contains all $1{,}024$ backgrounds. Local methods can sample only the empty background and ten singleton backgrounds. At each phase, the active method scores measurements by their root mean square sensitivity over unresolved contrasts, scaled by the current distance of those contrasts from zero, and mixes in a two percent uniform allocation. The allocation is frozen before fresh samples are drawn, as required by Theorem 5.

The safe regime has minimum true gap $.07$, the unsafe regime has minimum true gap $-.10$, and the near boundary regime has minimum true gap $.01$. Each method and regime uses 100 independent

repetitions. There are no incorrect definite answers among the 900 evaluations. The absence of observed errors is descriptive. The statistical guarantee comes from Theorem 5 rather than from the finite empirical count.

E Additional Figures

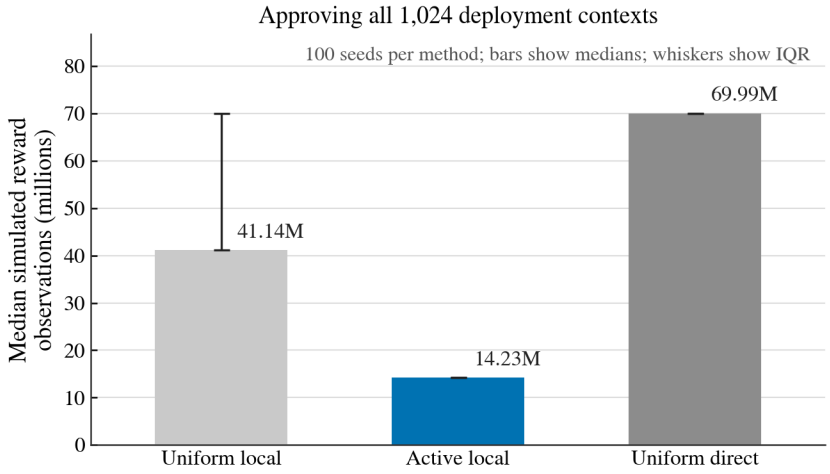

Figure 3: Simultaneous approval in the safe regime. Whiskers are interquartile ranges across 100 repetitions. All methods approve all 100 instances.

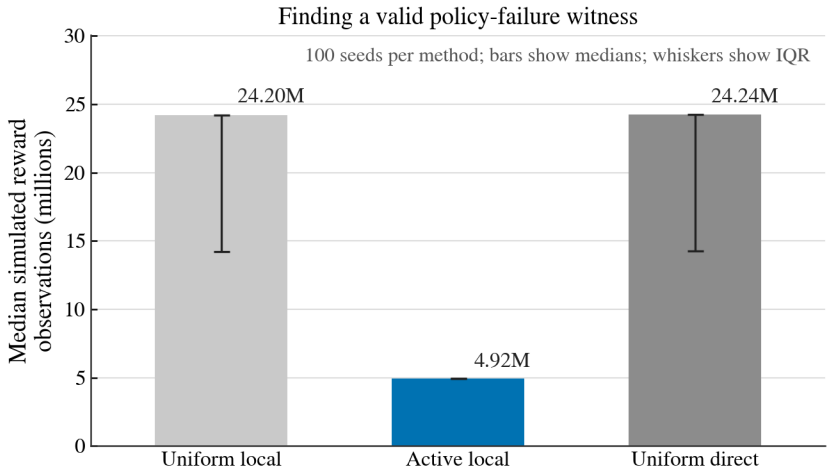

Figure 4: Refutation in the unsafe regime. All methods return a valid negative witness in all 100 instances.

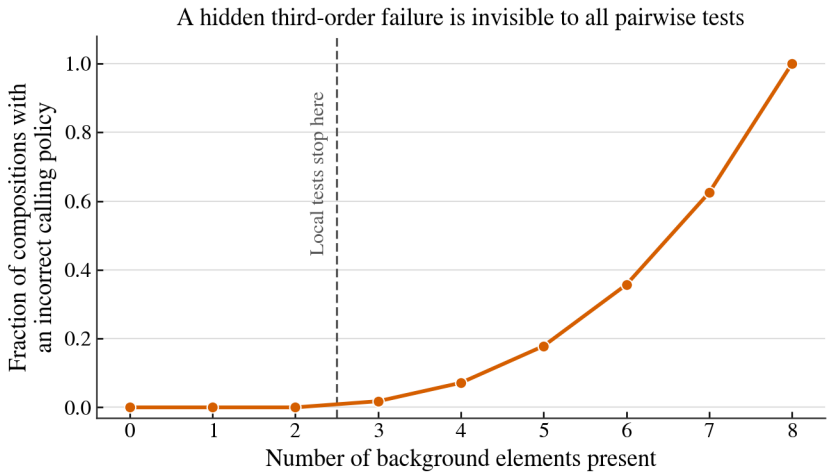

Figure 5: Hidden third order interactions are invisible to all quadratic local tests but create failing deployment backgrounds. This experiment verifies the information obstruction rather than a learned procedure for discovering the hidden gate.