CERTIFYING THE INFORMATION VALUE OF ADAPTIVE TOOL USE

Anonymous authors
Paper under double-blind review

ABSTRACT

An early network alarm can guide a tool-using agent's next query. Yet when an interactive agent outperforms one that chooses all queries in advance, the gain may reflect better information, easier sequential processing, or additional inference compute. We define adaptive information value as the improvement in optimal expected task utility from choosing queries based on previous responses rather than fixing them in advance, under the same task and query budget. We derive a sound finite-budget certificate that upper-bounds optimal adaptive performance without enumerating adaptive policies. The certificate combines pairwise response-channel overlap with utility-sensitive fractional matching and supports arbitrary finite priors and utilities, heterogeneous query costs, multivalued responses, and memoryless noise. Controlled network-diagnosis experiments show that changing only the tool interface can create or eliminate adaptive information value. Across 63 selected topology lineages and three language-model families, externally optimized adaptive evidence improves prediction accuracy by 5.84 percentage points on the positive-gap interface, with a 4.37-point larger gain than on a matched zero-gap control; both effects remain significant after Holm correction. However, when the models select their own queries, adaptive querying underperforms pre-committed querying in our tests, distinguishing the ability to use adaptive evidence from the ability to select it. Our framework provides a benchmark-level audit of whether feedback creates an information advantage before attributing multi-turn gains to adaptive acquisition.

1 INTRODUCTION

Tool-using AI agents are designed for tasks in which useful evidence must be acquired sequentially, from network diagnosis and troubleshooting to knowledge elicitation. Their central promise is not merely that they can make several tool calls, but that each response can guide what they ask next. An initial diagnostic, for example, may localize a failure to one part of a network, allowing the next query to target the remaining uncertainty rather than follow a fixed script. This feedback loop motivates benchmarks that compare interactive agents with one-shot or pre-committed baselines (Lu et al., 2025; Yao et al., 2025; Guo et al., 2026). Yet a higher multi-turn score alone does not establish that adaptive query selection provided more useful information.

As a schematic network-diagnosis example, suppose a communication network has four labeled connections (graph edges), A–D, exactly one of which has failed, and only two diagnostic calls are allowed. The tool menu contains a regional scan that reports whether the fault lies in $\{A, B\}$ or $\{C, D\}$, a test that distinguishes A from B, and a test that distinguishes C from D. An adaptive policy scans first and then selects the relevant local test, identifying all four faults. Any policy that must choose two of these three tests in advance leaves at least one pair indistinguishable. Figure 1 shows this opportunity, why an observed interactive score gain still does not identify it, and how our audit resolves the ambiguity.

The observed advantage can arise through three mechanisms. Early answers may change which later evidence is most useful, creating an *information advantage*. A bounded model may instead process the same evidence more effectively when it is presented sequentially, creating a *presentation or optimization advantage*. Finally, a multi-turn protocol may provide more tool calls, tokens, or inference time, creating a *compute advantage* (Kapoor et al., 2025; Snell et al., 2025; Chen et al.,

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.

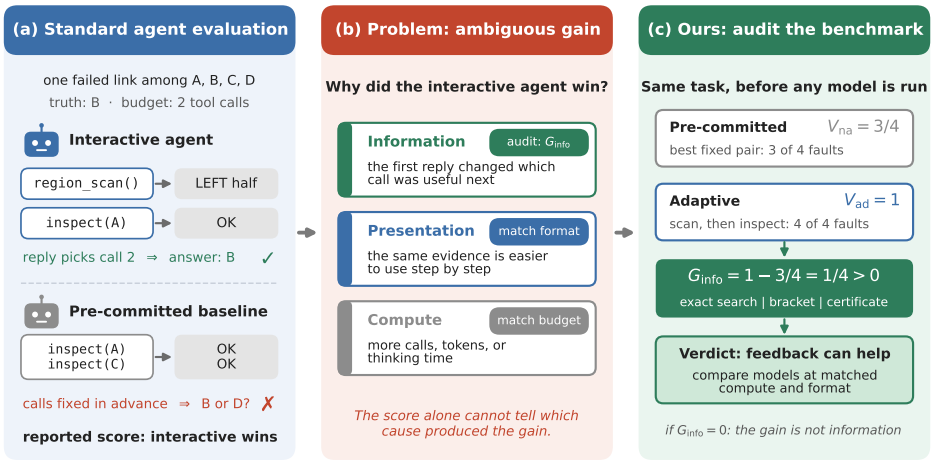

Figure 1: (a) With connection B as the hidden failed edge, the interactive agent's first reply selects a useful second call, whereas the pre-committed calls cannot separate B from D. (b) The score gain is consistent with three mechanisms; only information shows feedback helped. (c) Before any model runs, our audit computes the best pre-committed and adaptive values on this task; here $G_{\text{info}} = 1/4 > 0$, so feedback can help, and a model comparison must match compute and presentation.

2024b). Only the first identifies an information gain from feedback-conditioned acquisition. Without separating them, a multi-turn accuracy gain is an ambiguous measure of adaptivity.

We therefore study adaptivity at the level of the benchmark specification, before choosing or evaluating a model. The specification comprises a finite latent hypothesis space, a prior, a terminal utility, query costs and budget, and the response channel induced by each tool. We define the *adaptive information value* as the difference between the optimal Bayes value (Berger, 1985) of a policy that conditions each query on previous responses and that of a policy that must commit to its queries in advance. A zero value means that feedback creates no additional Bayes-optimal opportunity; it does not imply that the two protocols are equally easy for a bounded model.

Computing this value directly is difficult: deciding whether k fixed queries separate every hypothesis contains Minimum Test Set/Test Cover as a special case and is NP-complete (Garey & Johnson, 1979), while constructing optimal adaptive decision trees is also NP-complete in classical settings (Hyafil & Rivest, 1976). We derive a sound finite-budget certificate that upper-bounds adaptive Bayes value without enumerating adaptive policies. The certificate combines pairwise channel affinity, which controls how confusable two hypotheses must remain, with a utility-sensitive fractional matching that aggregates unavoidable decision regret without double counting. It supports nonuniform priors, arbitrary finite terminal utilities, heterogeneous query costs, multivalued responses, randomized stopping, repeated queries, and memoryless noise.

Our experiments test both whether a benchmark contains an adaptive opportunity and whether models can use it. The two questions are complementary: a positive value shows only that feedback *can* help, not that a given model exploits it, while a model gain on a zero-value benchmark cannot come from adaptive information. Controlled network-diagnosis experiments hold the latent problem fixed and show that changing only the tool interface can create or eliminate adaptive information value. Source-faithful audits extend the diagnostic beyond the constructed graph setting. Finally, a study across selected independent topology lineages and multiple language-model families finds that externally optimized adaptive evidence helps more on a positive-value interface than on a matched zero-value control. In practice, this means a tool-using agent, such as a network-diagnosis assistant, can benefit from feedback when its tool interface has a positive audited value and its queries are chosen by an explicit planner rather than left for the model to discover. Evidence-fixed and model-selected controls separate the ability to use adaptive evidence from presentation effects and from the harder problem of discovering an adaptive query policy. Our main contributions are:

- We define a benchmark-level estimand that isolates the information value of feedback-conditioned query selection from presentation, optimization, and compute effects.
- We prove a general finite-budget certificate for adaptive Bayes value that avoids policy enumeration and accommodates finite priors, utilities, costs, multivalued responses, and memoryless noise. It reduces to the classical two-point bound in the simplest case, and we show that class-level ratio guarantees cannot determine the gap of an individual specification.
- We show through controlled and source-faithful studies that adaptive information value depends on the tool interface, even when the underlying decision problem is unchanged.
- We show that language models can use externally optimized adaptive evidence where the specification provides an opportunity, while distinguishing evidence use from autonomous query-policy discovery.

2 Related Work

Active learning, generalized binary search, adaptive submodularity, and decision-region determination select informative observations (Dasgupta, 2004; Nowak, 2011; Golovin & Krause, 2011; Golovin et al., 2010; Javdani et al., 2014), and value-of-information and stochastic-probing work compares adaptive and fixed plans under priors, costs, and structured utilities (Krause & Guestrin, 2009; Gupta et al., 2017; Bradač et al., 2019). Batched bandits compare feedback schedules (Perchet et al., 2016; Gao et al., 2019), noisy twenty questions studies corrupted answers (Jedynak et al., 2012), amortized Bayesian design learns sequential policies, including decision-aware variants (Foster et al., 2021; Blau et al., 2022; Huang et al., 2024), and differentiable causal design studies non-adaptive batches (Tigas et al., 2023). We do not claim that adaptive design is new, nor do we provide an approximation guarantee for these objectives.

Clarification benchmarks and future-turn preference modeling evaluate or train language models to resolve missing information (Zhang et al., 2024; Li et al., 2025; Zhang et al., 2025), and multi-turn agent benchmarks evaluate interaction across tools and environments (Liu et al., 2024; Ma et al., 2024; Lu et al., 2025). These works optimize policies or measure interactive behavior; our target is the benchmark-level Bayes opportunity available before selecting a bounded agent.

The certificate's pairwise testing ingredient is classical in the Le Cam–Assouad tradition (Yu, 1997; Tsybakov, 2009), and controlled sensing and interactive lower bounds also study adaptive information collection (Nitinawarat et al., 2013; Naghshvar & Javidi, 2013; Kartik et al., 2022; Chen et al., 2024a). Our contribution is to make the adaptive-versus-pre-committed Bayes difference an explicit benchmark estimand and to combine a path-budget affinity floor, utility-sensitive regret, and fractional matching into an auditable finite-budget certificate; Section 4 states the precise separation (Table 1, Proposition 5).

3 Adaptive Information Value

Consider a finite latent-state decision problem with a nonempty hypothesis set H and prior p. After acquiring evidence, a policy selects an action from a finite nonempty set $\mathcal{Y}$ and receives finite-valued utility $u(a, h)$ when the true hypothesis is h. Each query q in a finite nonempty set $\mathcal{Q}$ has cost $c_q > 0$ and returns an observation z from a finite alphabet through a response channel $W_q(z \mid h)$; responses are conditionally independent given the hypothesis and selected queries, and every realized query path must respect a budget $C \geq 0$. An adaptive policy may choose each query after observing previous responses, whereas a pre-committed policy must choose its cost-feasible query collection in advance. In our network-diagnosis experiments, h identifies the failed connection, a query selects a node, and the response is determined by the diagnostic interface. This formulation keeps the underlying decision problem separate from the tool interface, allowing us to change what a query reveals while holding the hypothesis space, prior, utility, costs, and budget fixed.

Definition 1 (Adaptive information value)**.** Let $V_{\mathrm{na}}(C)$ be the maximum expected utility over policies that pre-commit to a finite, possibly repeated, cost-feasible query sequence, including the empty sequence, and then choose a terminal action from the resulting observations. Randomization over pre-committed sequences is allowed but unnecessary because a finite mixture cannot exceed its best

deterministic component. Let $V_{\mathrm{ad}}(C)$ be the maximum expected utility over non-anticipating policies that may randomize, stop, and repeat queries, with total cost at most C on every realized path. The *adaptive information value* is

$$G_{\mathrm{info}}(C) = V_{\mathrm{ad}}(C) - V_{\mathrm{na}}(C). \tag{1}$$

Because every pre-committed policy is a valid adaptive policy, $G_{\mathrm{info}}(C) \geq 0$.

Equation 1 is a property of the benchmark specification, not of a particular model implementation.

Proposition 2 (Noiseless identification special case)**.** *Assume a uniform prior, a deterministic query family F with unit costs, terminal action space $\mathcal{Y} = H$, and utility $u(a, h) = \mathbf{1}\{a = h\}$. Let $N(F, k)$ be the maximum number of nonempty answer-signature atoms induced by any k pre-committed queries from F, and let $A(F, k)$ be the maximum number of nonempty terminal cells in any depth-k adaptive decision tree over F. Then $V_{\mathrm{na}}(k) = N(F, k)/|H|$, $V_{\mathrm{ad}}(k) = A(F, k)/|H|$, and therefore $G_{\mathrm{info}}(k) = \big(A(F, k) - N(F, k)\big)/|H|$.*

The proof, a direct atom-counting argument, is in Appendix D.

Observed model contrasts. Fix a query budget C and a matched model-resource budget r. Let $m_{\mathrm{ad}}(C, r)$ and $m_{\mathrm{na}}(C, r)$ denote the expected utilities achieved by a model under the adaptive and pre-committed protocols, and define the optimization shortfalls $e_s(C, r) = V_s(C) - m_s(C, r)$ for $s \in \{\mathrm{ad}, \mathrm{na}\}$. Then

$$m_{\mathrm{ad}}(C, r) - m_{\mathrm{na}}(C, r) = G_{\mathrm{info}}(C) + e_{\mathrm{na}}(C, r) - e_{\mathrm{ad}}(C, r). \tag{2}$$

Thus, even under matched resources, an observed contrast mixes adaptive information value with how effectively the model uses each protocol, and unequal calls, tokens, or inference time add a compute-allocation term. This has two practical consequences for evaluation. First, a higher model score with feedback does not by itself establish an information gain from feedback, because the gain may come entirely from $e_{\mathrm{na}} - e_{\mathrm{ad}}$; $G_{\mathrm{info}}(C)$ must therefore be audited separately. Second, when $G_{\mathrm{info}}(C) = 0$, a nonzero model contrast is not a contradiction: it measures protocol optimization or compute, not adaptive information.

Compute confounding. Equal per-call limits do not give a valid information comparison: k calls receive k times the token allowance, so a synthetic model whose accuracy depends only on tokens gains up to $+0.327$ although $G_{\mathrm{info}} = 0$. On a certified-zero interface, the measured compute effect for Qwen3-8B takes either sign across budgets, so it cannot be assumed or subtracted as a fixed correction (Figure 4 and configuration in Appendix L). Evaluators should run the pre-committed condition under both equal-per-call and matched-total budgets and report measured token use.

4 A Finite-Budget Certificate

Computing $V_{\mathrm{ad}}(C)$ exactly requires searching over response-contingent decision trees. We instead upper-bound it by lower-bounding the regret every feasible adaptive policy must retain: within the budget, some pairs of hypotheses must stay partly confusable, and that unavoidable pairwise regret is aggregated without double counting.

For each hypothesis h, let $u^\star_h = \max_{a \in \mathcal{Y}} u(a, h)$ be its perfect-information utility. For a pair of hypotheses (i, j), define the minimum joint regret of using one terminal action for both by

$$\delta_{ij} = \min_{a \in \mathcal{Y}} \left[u^\star_i - u(a, i) + u^\star_j - u(a, j) \right].$$

For each query $q \in \mathcal{Q}$, define its Bhattacharyya affinity on the pair as

$$\rho^{ij}_q = \sum_z \sqrt{W_q(z \mid i) W_q(z \mid j)}.$$

Finally, define the smallest affinity product permitted by budget C:

$$\beta_{ij}(C) = \min_{\substack{T \geq 0,\ q_1, \ldots, q_T \in \mathcal{Q} \\ \sum_{t=1}^{T} c_{q_t} \leq C}} \prod_{t=1}^{T} \rho^{ij}_{q_t},$$

where the empty product is one. Every feasible adaptive policy induces transcript distributions under i and j whose affinity is at least $\beta_{ij}(C)$.

Theorem 3 (Finite-budget adaptive-value certificate)**.** *Let $p_i = p(i)$. For each pair $i < j$, define*

$$\omega_{ij} = \frac{p_i + p_j - \sqrt{(p_i + p_j)^2 - 4p_i p_j \beta_{ij}(C)^2}}{2}, \qquad d_{ij} = \delta_{ij}\omega_{ij}.$$

Let $L^\star$ be the fractional-matching value

$$L^\star = \max_{x_{ij} \geq 0} \sum_{i<j} x_{ij} d_{ij}$$

subject to

$$\sum_{j<i} x_{ji} + \sum_{j>i} x_{ij} \leq 1 \qquad \textit{for every } i \in H.$$

Then every non-anticipating adaptive policy that may randomize, stop, and repeat queries while respecting path budget C satisfies

$$V_{\mathrm{ad}}(C) \leq \sum_{h \in H} p(h) u_h^\star - L^\star. \tag{3}$$

Corollary 4 (Adaptive-information gap certificate)**.** *Combining the adaptive upper bound in Equation 3 with any certified lower bound $\underline{V}_{\mathrm{na}}(C) \leq V_{\mathrm{na}}(C)$,*

$$0 \leq G_{\mathrm{info}}(C) \leq \sum_{h \in H} p(h) u_h^\star - L^\star - \underline{V}_{\mathrm{na}}(C).$$

The proof (Appendix A) lower-bounds pairwise transcript affinity under every feasible adaptive policy, converts this affinity into prior-weighted overlap and unavoidable decision regret, and aggregates the pairwise regrets through fractional matching without double counting.

Relation to prior results. The ingredients are classical. Step 1 of the proof is an adaptive tensorization of Bhattacharyya affinity, and for $|H| = 2$ with 0–1 utility Theorem 3 reduces to the classical Bhattacharyya lower bound on two-hypothesis Bayes error, evaluated at the budgeted floor $\beta_{ij}(C)$. Prior risk inequalities can sometimes be specialized into instance-wise adaptive upper bounds—our Fano comparator below is one example—and exact value of information is available for special graphical structures. Our contribution is a directly computable, utility-sensitive finite-budget certificate for the stated general finite model, paired with a pre-committed lower bound and a deterministic cardinality bracket to return zero, positive, or inconclusive benchmark-audit verdicts (Table 1). Adaptivity-gap theorems instead bound worst-case ratios $V_{\mathrm{ad}}/V_{\mathrm{na}}$ over structured instance classes; the next result shows why such a ratio cannot determine the gap of one latent-state specification.

Proposition 5 (Class-level ratios do not determine instance value)**.** *Under the assumptions of Proposition 2, let $|H| = n = 2^k$. The family of dyadic intervals of H gives $V_{\mathrm{ad}}(k) = 1$ and $V_{\mathrm{na}}(k) \leq (k+1)/n$, so $V_{\mathrm{ad}}/V_{\mathrm{na}} \geq n/(\log_2 n + 1)$ is unbounded in n; the family 2^H gives $G_{\mathrm{info}}(k) = 0$ at the same n and k.*

The proof is in Appendix D. Generalized Bayes-risk inequalities extend Fano, Le Cam, and Assouad to nonuniform priors and broader losses (Chen et al., 2016); interactive versions address adaptive data collection (Chen et al., 2024a). Our implemented sharp-Fano comparator (Cover & Thomas, 2006) is the standard uniform-identification specialization. On the 36 validation cells where that comparator applies, neither bound dominates: the certificate is tighter in 14, equal in 8, and looser in 14 (mean slack 0.166 versus 0.137), and their minimum lowers mean slack to 0.112. That comparator is not defined for the other 72 cells, which have nonuniform priors or decision utilities; this is not a claim that generalized Fano methods are inapplicable.

Computation. For commensurate costs with scaled integer budget B, a one-dimensional recursion computes every affinity floor in $O(|H|^2 B|\mathcal{Q}|)$ time, followed by a fractional-matching linear program (Appendix B). When tighter bounds are needed, an exact two-hypothesis regret recursion can replace each affinity-based edge, and an anytime Bellman search can further tighten the adaptive

Table 1: Relationship to prior theory. A direct audit returns a zero, positive, or inconclusive verdict for one fully specified adaptive-versus-pre-committed problem without an additional reduction.

line of work	bound and scope	direct audit?
adaptivity gaps (Gupta et al., 2017; Bradač et al., 2019)	worst-case $V_{\mathrm{ad}}/V_{\mathrm{na}}$ over independent-item classes	no
Bayes-risk bounds (Yu, 1997; Chen et al., 2016; 2024a)	adaptive or passive risk under broad priors and losses	after specialization
active testing (Naghshvar & Javidi, 2013; Kartik et al., 2022)	typically asymptotic error exponents and policies	no
adaptive submodularity (Golovin & Krause, 2011; Nowak, 2011)	greedy-versus-optimal adaptive guarantees; lower bounds on V_{ad}	no
exact nonmyopic VOI (Krause & Guestrin, 2009)	optimal adaptive and fixed plans; efficient exact algorithms on chains	yes, structured
this work (Thm. 3, bracket)	instance-level upper bound on V_{ad} and lower bound on V_{na} for finite p, u, c, W	**zero, positive, or inconclusive**

upper bound; both may require exponential work. The core certificate is always sound but can be inconclusive, and an inconclusive bound must not be reported as a zero adaptive information value. History-dependent tool failures require augmenting the latent state and are not covered directly by the memoryless-channel formulation.

Deterministic channels. For deterministic tools, $\rho_q^{ij} \in \{0, 1\}$. If any affordable query separates i and j, then $\beta_{ij}(C) = 0$ and that pair contributes no regret to the matching bound. Consequently, the affinity certificate does not encode the separate cardinality constraint that a depth-k tree has at most the product of its response-alphabet sizes in leaves. This is a structural source of looseness, not a soundness failure. Under the assumptions of Proposition 2 (deterministic responses, uniform prior, exact identification), a noiseless bracket supplies this missing atom-count constraint by bounding both capacities directly: constructive designs and decision trees give lower bounds N_{lb} and A_{lb}, and pair-coverage submodularity and a leaf-capped Bellman recursion give upper bounds N_{ub} and A_{ub} (Appendix C). Hence $(A_{\mathrm{lb}} - N_{\mathrm{ub}})/|H| \le G_{\mathrm{info}} \le (A_{\mathrm{ub}} - N_{\mathrm{lb}})/|H|$, which certifies a zero when $A_{\mathrm{ub}} = N_{\mathrm{lb}}$ and a positive gap when $A_{\mathrm{lb}} > N_{\mathrm{ub}}$.

Validation. We compare the certificate with exact adaptive values on 108 synthetic instances spanning 4–10 hypotheses, multivalued noisy channels, three prior-and-utility regimes, and budgets of two or three. No bound is violated; mean slack falls from 0.223 for the perfect-information bound to 0.166 for affinity and 0.135 for the exact-pair tightening, and the best tightening is exact in 37.0% of cells (Table 5, Appendix B). These checks validate the implementation rather than prove the theorem.

On a predeclared panel of 72 real-topology interface–budget cells, the scalable affinity certificate certifies 15 zero gaps and leaves 57 inconclusive, while the noiseless bracket classifies 35 cells as zero, 13 as strictly positive, and 24 as inconclusive. All 15 affinity-certified zeros are saturation cells: the pre-committed lower bound already reaches perfect-information utility. In a separate operator-motivated sensitivity scenario (nonuniform prior, asymmetric utility, heterogeneous costs, noisy responses; not measured operator data), the certificate strictly improves on the perfect-information bound in all 24 cells, with median runtime 0.015 seconds, but resolves none as zero or positive; failure to certify a zero is not evidence of a positive gap. Accordingly, the controlled results (Tables 2–3, Figure 2, and the noise experiments) use completed exact searches, not Theorem 3; the uncapped-menu results use the bracket; and the 108-cell validation and 24-cell sensitivity use the affinity certificate. Exact search is strongest at small scale, the bracket handles deterministic atom-count constraints, and affinity is the general noisy-channel fallback.

5 Controlled Interface Experiments

We test whether adaptive information value changes when only the tool response changes. The hidden hypothesis is a single failed bridge (an edge whose removal disconnects the graph) in a

Table 2: Changing only the response changes adaptive information value (exact results; 100 constructed single-bridge instances; budget three).

interface	instances with $G_{\mathrm{info}} > 0$	mean G_{info}	response
neighbor readout	0/100	0.0000	structural detail
alarm-count readout	97/100	**0.1607**	coarse aggregate

Table 3: Mean exact-identification accuracy on 25 real topologies. Values are rounded separately, so printed differences may vary by 0.001.

	neighbor readout		alarm-count readout	
method	$k = 2$	$k = 3$	$k = 2$	$k = 3$
random fixed	0.229	0.308	0.194	0.257
greedy fixed	0.652	0.809	0.242	0.322
optimal fixed	0.652	0.809	0.246	0.326
myopic adaptive	0.652	0.809	0.288	0.459
optimal adaptive	0.652	0.809	0.313	0.483
information gap	**0.000**	**0.000**	**0.067**	**0.156**

known graph, the prior is uniform, the terminal utility is exact identification, and each query selects a node under a budget of three calls. We compare a structural neighbor readout with a coarse alarm-count readout while holding the graph, hypothesis space, prior, utility, candidate-query menu, and budget fixed. The primary panel contains 100 constructed graphs from five random-graph families, each formed from a 20-node core with the same 12 seeded pendant faults. We report $N(F, k)$ and $A(F, k)$ as exact only when both optimization procedures certify completion; otherwise, we report explicit bounds.

Changing only the response interface changes adaptive information value (Table 2). The structural interface has exact zero gap on all 100 constructed graphs, whereas the local aggregate interface has a positive gap on 97, with mean 0.1607. These are controlled mechanism probes, not prevalence estimates over naturally occurring networks.

The neighbor-readout zero does not arise because the task is already solved. On 25 real topologies at budget three, optimal fixed and adaptive accuracies are both 0.809, while random fixed querying reaches only 0.308. Under the alarm-count readout, optimal accuracy rises from 0.326 with fixed queries to 0.483 with adaptive queries, leaving substantial adaptive headroom (Table 3). Greedy fixed and myopic adaptive baselines also coincide on the neighbor readout at both budgets, an algorithm-dependent diagnostic rather than a replacement for the Bayes-optimal estimand.

Figure 2 separates aggregation from truncation. Local aggregate responses can preserve uncertainty in a form that makes the next best query depend on earlier answers, whereas structural or global responses often allow pre-committed queries to be equally informative. Shortening a rich response instead makes the task harder without creating adaptive value: reducing the row cap from 500 to one lowers the optimal accuracy ceiling from 0.813 to 0.288 (0.239 with no rows), yet the gap stays exactly zero at every tested cap. The mechanism is thus the partition of hypotheses induced by each response, not output length; the shipped cap binds on none of the audited graphs, so this is a mechanistic intervention rather than a claim about deployed load.

Robustness and transfer. Adaptive information value is a property of the complete response channel, not the interface name: symmetric noise shrinks the alarm-count gap from 0.0729 to 0.0050 at error 0.50 while the neighbor-readout zero persists through error 0.30, but asymmetric false-alarm, miss, and erasure channels give most tested topologies small positive gaps (Appendix F). On uncapped menus beyond exact enumeration, the bracket certifies zero gap on every tested structural network at budgets two and three, and on 83% and 80% of networks for full three-family menus; it loosens beyond budget three, so we make no general large-budget zero-gap claim. On the pinned TwentyQuestions matrix from Adaptive Elicitation (Wang et al., 2025), with an explicitly custom prior and utility, the mean exact gap reaches 0.0479 after four follow-up questions; this is a source-faithful oracle counterfactual, not an evaluation of the released learned policy (Appendix H).

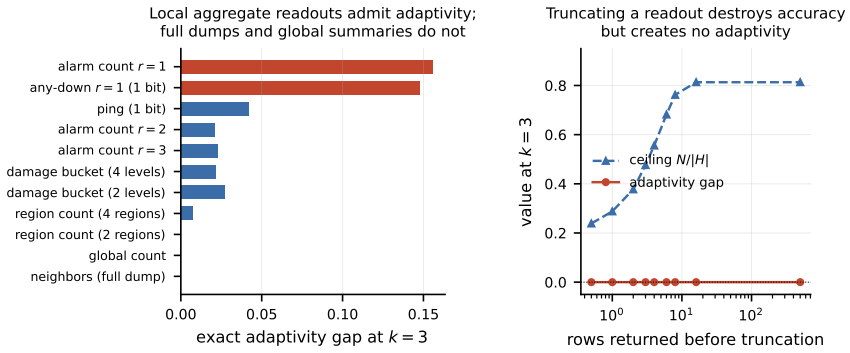

Figure 2: Left: exact adaptive information gaps across response interfaces on 25 real topologies, budget three. Right: deterministic truncation lowers achievable accuracy but does not create a gap.

Table 4: Model confirmation: paired accuracy differences averaged over three model families.

endpoint	effect	95% CI	Holm-adjusted p
alarm-count contrast	**+0.0584**	**[+0.0373,+0.0800]**	$\mathbf{3.0\times10^{-5}}$
interface interaction	**+0.0437**	**[+0.0190,+0.0690]**	**0.00282**
neighbor equivalence (TOST, ±0.03)	+0.0148	[+0.0009,+0.0289]	0.0194

6 Can Models Use the Certified Opportunity?

A positive specification-level gap shows that better adaptive evidence is available, but not that a language model can use it. We test this distinction in a prospectively specified study, with self-attested timing and disclosed deviations, on 63 independent topology lineages (distinct operator networks, with snapshots of one operator merged) selected to have a positive exact alarm-count gap and a zero exact neighbor-readout gap. The study evaluates Qwen2.5-32B, Mistral-Small-24B, and Llama-3.3-70B with two decoding seeds and 24 paired truth draws per lineage. For the primary comparison, an external exact optimizer generates either an optimal adaptive transcript or an optimal pre-committed transcript, and both are presented in the same final-prediction format. This design isolates the model's ability to use adaptively selected evidence from the separate ability to discover an adaptive query policy.

Confirmatory analysis. The primary endpoint is the paired adaptive-minus-pre-committed accuracy difference on the alarm-count readout. The second endpoint is the interaction between this contrast and the corresponding contrast on the neighbor readout. The third tests whether the neighbor-readout contrast is equivalent to zero within the prospectively chosen margin of ±0.03. Observations are averaged within each model and lineage, so the lineage is the independent unit; tests use lineage-level sign reversals, equivalence uses two one-sided tests, and Holm correction is applied to the three confirmatory p-values (Appendix I).

Table 4 shows that externally optimized adaptive evidence improves alarm-count accuracy by 0.0584 (rounded raw accuracies 0.249 and 0.308), and the effect is positive in every model family and every leave-one-family-out estimate (Appendix I). On the zero-gap neighbor readout, the effect is 0.0148 and satisfies the prospectively chosen ±0.03 equivalence criterion; equivalence is margin-dependent and does not establish a literal zero effect. The positive interaction shows that the benefit is materially larger where the specification provides an adaptive information opportunity. Because the lineages were selected for this mechanism contrast and the policies were optimized externally, the result demonstrates that the tested models can use externally optimized adaptive evidence in the selected regime, not prevalence across tasks or autonomous adaptive-policy discovery.

The interaction is the paired difference between the two interface contrasts, not an independent replication. Holm correction applies only to the three endpoints in Table 4; the controls below are outside that family, and a post hoc studentized check is in Appendix I.

Controls. Revealing the same fixed evidence sequentially rather than jointly changes accuracy by -0.0123 (95% CI $[-0.0260, 0.0009]$), so the measured presentation change does not account for the positive primary effect. In the model-selected comparison, adaptive querying performs worse than pre-committed querying by 0.0324 (95% CI $[-0.0531, -0.0141]$), even though the same models benefit when adaptive evidence is selected externally. Together, these controls distinguish two capabilities: using informative evidence and discovering an effective adaptive query policy, with policy discovery remaining a major measured bottleneck in this regime.

The positive neighbor-readout contrast does not contradict $G_{\text{info}} = 0$: by Equation 2, it reflects optimization shortfalls and sampling variation, since the fixed-evidence control isolates reveal schedule but not the full optimization problem. In all four primary condition-by-interface cells, exact-format parse rate is 1.000, final-answer truncation is zero, and mean token use differs by at most 2.5% between conditions, ruling out parsing and truncation asymmetries and sharply constraining token-opportunity imbalance relative to simpler multi-turn comparisons (Appendix I; one qualification deviation is disclosed in Appendix J).

Negative and inconclusive results. Both tested reasoning models failed the reportability gate in every sweep cell, so those cells are unmeasurable rather than negative capability results (Appendix L); a prospectively specified five-interface follow-up was inconclusive for a monotone relation between gap magnitude and model benefit (Appendix J); and no frontier or non-quantized model was tested.

7 A Practical Evaluation Protocol

The framework suggests five steps for evaluating claims about adaptive tool use.

1. **Specify the problem.** Declare the hypotheses, prior, actions and utility, response channels, query costs, and path budget; a tool name alone does not define an information-acquisition problem.
2. **Bound the opportunity.** Lower-bound V_{na}, upper-bound V_{ad} with the tightest valid bound (Section 4), and report the interval for G_{info}, giving no verdict if it is inconclusive.
3. **Power the study.** The oracle gap is a maximum effect, so a null result is uninformative below the detectable effect (Appendix N).
4. **Match resources and presentation.** Match total resources, record token use, truncation, and parsing, present both evidence types in one format, and add a joint-versus-sequential control.
5. **Label the scope.** Call a result native only when the original distribution, utility, interface, budget, and implementation are kept, and otherwise a source-faithful counterfactual or proxy.

8 Limitations and Conclusion

The exact headline experiments primarily use single-failure hypothesis spaces and small query budgets, and the certificate can become loose at larger budgets. The controlled graphs are mechanism probes rather than prevalence samples. The model study uses selected topology lineages, three quantized open-weight instruction-tuned families, and externally optimized policies; it establishes that these models can use improved adaptive evidence in the selected regime, not that they discover effective policies autonomously, that the effect generalizes to frontier models, or that it reflects production network diagnosis. We do not provide a native audit of ToolSandbox or τ-bench, and the GDS-Agent audit (Appendix H) is schema-level, without live Neo4j execution. Applying the certificate natively requires a finite, stable latent state with an explicit prior, terminal utility, response channel, costs, and budget.

A benchmark can remain difficult and useful when $G_{\text{info}} = 0$, but its multi-turn score then measures orchestration, computation, or bounded-model optimization rather than the information value of conditioning later queries on earlier answers. Before attributing a multi-turn gain to adaptive acquisition, evaluators should bound both Bayes values and control for compute and presentation.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing and reorganization, language polishing, and reviewing author-written code for potential bugs and debugging issues. The authors wrote the research code and independently reviewed and tested all AI-suggested corrections. We did not use generative AI tools to generate data or model outputs. The authors independently verified all mathematical statements and proofs and take full responsibility for the final content of this work, including its text, claims, code, and figures.

ETHICS STATEMENT

This work uses public network-topology and benchmark data, involves no human subjects, and contains no personal information. The primary risk of misinterpretation is overgeneralization: interface-level mechanism results could be misread as claims that multi-turn agents, particular benchmarks, or production network-diagnosis systems are broadly ineffective. We mitigate this risk by distinguishing native, source-faithful, and proxy evidence; reporting zero, positive, and inconclusive verdicts; and stating the limits of the selected tasks, budgets, and model roster. We make no claims about deployed systems or proprietary frontier models.

REPRODUCIBILITY STATEMENT

An anonymized artifact is provided as supplementary material and retains code, deterministic search-completion flags, dataset and model metadata, serving versions, qualification records, machine-readable results, raw transcripts, and source manifests. After fetching the pinned Topology Zoo snapshot with "python3 scripts/fetch_data.py", the command "python3 src/reproduce.py" reruns theorem, protocol, manuscript-structure, and claim-to-artifact checks. The Topology Zoo mirror is pinned at commit `51118849` and covered by a 282-file SHA-256 manifest; Adaptive Elicitation code and data are pinned at commits `8bcda3b1` and `1cf8c2f2`, respectively. Proofs appear in Appendices A–D, processing details in Appendices E and I, the claim-scope crosswalk in Appendix G, and deviations and corrections in Appendices J and O. Historical byte-level snapshots of the confirmatory generation runner are unavailable, so we do not claim byte-for-byte regeneration; retained outputs can nevertheless be rescored and the lineage analysis rerun. Because version control began after the reported computations, prospective timing is documented by author attestation rather than commit history. An independent implementation found no combinatorial disagreement across approximately 61,000 differential cases, including exact-rational noisy checks; Appendix O records the reporting and implementation issues identified during this validation and their corrections.

REFERENCES

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, 2 edition, 1985. doi: 10.1007/978-1-4757-4286-2.

Tom Blau, Edwin V. Bonilla, Iadine Chades, and Amir Dezfouli. Optimizing sequential experimental design with deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2107–2128. PMLR, 2022. URL https://proceedings.mlr.press/v162/blau22a.html.

Domagoj Bradač, Sahil Singla, and Goran Žužić. (Near) optimal adaptivity gaps for stochastic multi-value probing. In *APPROX/RANDOM*, volume 145, pp. 49:1–49:21, 2019. doi: 10.4230/LIPIcs.APPROX-RANDOM.2019.49.

Fan Chen, Dylan J. Foster, Yanjun Han, Jian Qian, Alexander Rakhlin, and Yunbei Xu. Assouad, Fano, and Le Cam with interaction: A unifying lower bound framework and characterization for bandit learnability. In *Advances in Neural Information Processing Systems*, volume 37, pp. 75585–75641, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8a23a95e26d016711c0d70f79ade3c95-Abstract-Conference.html.

Lingjiao Chen, Jared Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LM calls all you need? Towards the scaling properties of compound AI sys-

tems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 45767–45790, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/51173cf34c5faac9796a47dc2fdd3a71-Abstract-Conference.html.

Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On Bayes risk lower bounds. *Journal of Machine Learning Research*, 17(218):1–58, 2016. URL https://jmlr.org/papers/v17/16-185.html.

Robert J. Connor. Sample size for testing differences in proportions for the paired-sample design. *Biometrics*, 43(1):207–211, 1987. doi: 10.2307/2531961.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. doi: 10.1002/047174882X.

Peter Damaschke. Combinatorial search in two and more rounds. *Theoretical Computer Science*, 780:1–11, 2019. doi: 10.1016/j.tcs.2019.02.004.

Sanjoy Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems*, volume 17, pp. 337–344, 2004. URL https://proceedings.neurips.cc/paper/2004/hash/c61fbef63df5ff317aecdc3670094472-Abstract.html.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3384–3395. PMLR, 2021. URL https://proceedings.mlr.press/v139/foster21a.html.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems*, volume 32, pp. 501–511, 2019.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

Dániel Gerbner and Máté Vizer. Rounds in a combinatorial search problem. *Discrete Applied Mathematics*, 276:60–68, 2020. doi: 10.1016/j.dam.2019.09.019. URL https://arxiv.org/abs/1611.10133.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011. doi: 10.1613/jair.3278.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal Bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, volume 23, pp. 766–774, 2010. URL https://proceedings.neurips.cc/paper/2010/hash/1e6e0a04d20f50967c64dac2d639a577-Abstract.html.

Xiaojun Guo, Mingxue Tian, Chenheng Zhang, Xiaohan Wang, Jiajun Chai, Guojun Yin, Wei Lin, Yifei Wang, and Yisen Wang. GRASP: Graph reasoning via agentic solving and probing of LLMs. In *International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/61718. Code: https://github.com/PKU-ML/GRASP.

Anupam Gupta, Viswanath Nagarajan, and Sahil Singla. Adaptivity gaps for stochastic probing: Submodular and XOS functions. In *ACM-SIAM Symposium on Discrete Algorithms*, pp. 1688–1702, 2017. doi: 10.1137/1.9781611974782.111.

Daolang Huang, Yujia Guo, Luigi Acerbi, and Samuel Kaski. Amortized Bayesian experimental design for decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109460–109486, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c59f05d7ab3638b138cc61f32e1a7cd1-Abstract.html.

Laurent Hyafil and Ronald L. Rivest. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1):15–17, 1976. doi: 10.1016/0020-0190(76)90095-8.

Shervin Javdani, Yuxin Chen, Amin Karbasi, Andreas Krause, Drew Bagnell, and Siddhartha S. Srinivasa. Near optimal Bayesian active learning for decision making. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 430–438. PMLR, 2014. URL https://proceedings.mlr.press/v33/javdani14.html.

Bruno Jedynak, Peter I. Frazier, and Raphael Sznitman. Twenty questions with noise: Bayes optimal policies for entropy loss. *Journal of Applied Probability*, 49(1):114–136, 2012. doi: 10.1239/jap/1331216837.

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.

Dhruva Kartik, Ashutosh Nayyar, and Urbashi Mitra. Fixed-horizon active hypothesis testing. *IEEE Transactions on Automatic Control*, 67(4):1882–1897, 2022. doi: 10.1109/TAC.2021.3090742.

Gyula O. H. Katona. Combinatorial search problems. In *A Survey of Combinatorial Theory*, pp. 285–308. North-Holland, 1973. doi: 10.1016/B978-0-7204-2262-7.50028-4.

Simon Knight, Hung X. Nguyen, Nick Falkner, Rhys Bowden, and Matthew Roughan. The Internet Topology Zoo. *IEEE Journal on Selected Areas in Communications*, 29(9):1765–1775, 2011. doi: 10.1109/JSAC.2011.111002.

Andreas Krause and Carlos Guestrin. Optimal value of information in graphical models. *Journal of Artificial Intelligence Research*, 35:557–591, 2009. doi: 10.1613/jair.2709.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *ACM Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Belinda Z. Li, Been Kim, and Zi Wang. QuestBench: Can LLMs ask the right question to acquire information in reasoning tasks? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c42c8d51556fabb4b57fc86d3d3d0d09-Abstract-Datasets_and_Benchmarks_Track.html.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.65. URL https://aclanthology.org/2025.findings-naacl.65/.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. AgentBoard: An analytical evaluation board of multi-turn LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/877b40688e330a0e2a3fc24084208dfa-Abstract-Datasets_and_Benchmarks_Track.html.

Olli S. Miettinen. The matched pairs design in the case of all-or-none responses. *Biometrics*, 24(2):339–352, 1968. doi: 10.2307/2528039.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013. doi: 10.1214/13-AOS1144.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978. doi: 10.1007/BF01588971.

Sirin Nitinawarat, George K. Atia, and Venugopal V. Veeravalli. Controlled sensing for multi-hypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013. doi: 10.1109/TAC.2013.2261188.

Robert D. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906, 2011. doi: 10.1109/TIT.2011.2169298.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016. doi: 10.1214/15-AOS1381.

Borun Shi and Ioannis Panagiotas. GDS agent for graph algorithmic reasoning. *arXiv preprint arXiv:2508.20637*, 2025. URL https://arxiv.org/abs/2508.20637.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Panagiotis Tigas, Yashas Annadani, Desi R. Ivanova, Andrew Jesson, Yarin Gal, Adam Foster, and Stefan Bauer. Differentiable multi-target causal Bayesian experimental design. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34263–34279. PMLR, 2023. URL https://proceedings.mlr.press/v202/tigas23a.html.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. doi: 10.1007/b13794.

Jimmy Wang, Thomas P. Zollo, Richard Zemel, and Hongseok Namkoong. Adaptive elicitation of latent information using natural language. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65897–65917. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25ff.html.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Bin Yu. Assouad, Fano, and Le Cam. In David Pollard, Erik Torgersen, and Grace L. Yang (eds.), *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997. doi: 10.1007/978-1-4612-1880-7_29.

Michael J. Q. Zhang, W. Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach LLMs to ask clarifying questions. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cwuSAR7EKd.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10746–10766. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.578. URL https://aclanthology.org/2024.acl-long.578/.

A Proof of the Finite-Budget Certificate

Step 1: transcript affinity. Fix a feasible adaptive policy and hypotheses i and j. Let P_i and P_j denote the laws of the finite observable transcript, including selected queries, observed responses, and stopping, after integrating out private policy randomness. At any shared transcript prefix, non-anticipation requires the behavioral policy to use the same finite conditional distribution over its next action under both hypotheses. If it selects query q, the affinity between the two conditional response distributions is ρ_q^{ij}.

Let $c_{\min} = \min_{q \in \mathcal{Q}} c_q > 0$. Any path with remaining budget b contains at most $\lfloor b/c_{\min} \rfloor$ further queries, and the definition of β_{ij} obeys

$$\beta_{ij}(b) = \min\left\{1, \min_{q:\, c_q \le b} \rho_q^{ij}\, \beta_{ij}(b - c_q)\right\}.$$

Backward induction on the maximum remaining path length, with stopping as the empty-product case, therefore yields

$$\sum_{\tau} \sqrt{P_i(\tau) P_j(\tau)} \ge \beta_{ij}(C).$$

Randomized stopping and query choice are covered by averaging over this common conditional action distribution at every shared history.

Step 2: affinity implies prior-weighted overlap. Define the weighted transcript measures

$$\mu_i(\tau) = p_i P_i(\tau), \qquad \mu_j(\tau) = p_j P_j(\tau),$$

and their total overlap

$$M_{ij} = \sum_{\tau} \min\{\mu_i(\tau), \mu_j(\tau)\}.$$

The weighted affinity satisfies

$$\sum_{\tau} \sqrt{\mu_i(\tau)\mu_j(\tau)} \ge \sqrt{p_i p_j}\, \beta_{ij}(C).$$

For each transcript, write the two masses as their minimum and maximum. Cauchy–Schwarz then gives

$$\left(\sum_{\tau} \sqrt{\mu_i(\tau)\mu_j(\tau)}\right)^2 \le M_{ij}\big(p_i + p_j - M_{ij}\big).$$

Because $M_{ij} \le \min\{p_i, p_j\} \le (p_i + p_j)/2$, solving the resulting quadratic inequality yields

$$M_{ij} \ge \frac{p_i + p_j - \sqrt{(p_i + p_j)^2 - 4 p_i p_j \beta_{ij}(C)^2}}{2} = \omega_{ij}.$$

Step 3: utility regret and fractional matching. Let $a(\tau)$ be the terminal action selected after transcript τ. For every pair (i, j) and transcript τ,

$$\begin{aligned}\mu_i(\tau)\big[u_i^\star - u(a(\tau), i)\big] + \mu_j(\tau)\big[u_j^\star - u(a(\tau), j)\big] \\ \ge \delta_{ij} \min\{\mu_i(\tau), \mu_j(\tau)\}.\end{aligned}$$

Summing over transcripts and applying Step 2 shows that the combined regret of hypotheses i and j is at least

$$\delta_{ij} M_{ij} \ge \delta_{ij} \omega_{ij} = d_{ij}.$$

Now multiply each pairwise inequality by a feasible fractional-matching weight x_{ij} and sum over pairs. Because

$$\sum_{j<i} x_{ji} + \sum_{j>i} x_{ij} \le 1,$$

the regret associated with any hypothesis is counted with total coefficient at most one. The policy's total expected regret is therefore at least

$$\sum_{i<j} x_{ij} d_{ij}.$$

Maximizing over feasible weights gives total regret at least $L^\star$. Subtracting this quantity from the perfect-information value yields

$$V_{\mathrm{ad}}(C) \le \sum_{h \in H} p(h) u_h^\star - L^\star,$$

which proves Theorem 3. Combining this upper bound with any certified lower bound on $V_{\mathrm{na}}(C)$ proves Corollary 4.

B Exact and Anytime Tightenings

Computing the affinity floor. For commensurate costs, choose $\Delta > 0$ such that $\bar{c}_q = c_q/\Delta$ and $B = C/\Delta$ are integers. The affinity floor is computed by the one-dimensional recursion

$$\bar{\beta}_{ij}(0) = 1, \qquad \bar{\beta}_{ij}(b) = \min\left\{1,\ \min_{q:\,\bar{c}_q \le b} \rho_q^{ij}\, \bar{\beta}_{ij}(b - \bar{c}_q)\right\}. \tag{4}$$

Then $\beta_{ij}(C) = \bar{\beta}_{ij}(B)$. Equation 4 is evaluated for $b = 0, \dots, B$. Here and below, a minimum over an empty feasible-query set is $+\infty$. Computing these values for all pairs costs $O(|H|^2 B |\mathcal{Q}|)$, followed by a fractional-matching linear program. The dependence on B is pseudo-polynomial in the scaled numerical budget, rather than polynomial in its binary encoding.

Exact-pair recursion. For a hypothesis pair (i, j) with unnormalized masses (x_i, x_j) and remaining budget b, define

$$\begin{aligned} R_{ij}(x_i, x_j, b) = \min\Bigg\{ & \min_{a \in \mathcal{Y}} \sum_{h \in \{i,j\}} x_h \big(u_h^\star - u(a,h)\big), \\ & \min_{q:\, c_q \le b} \sum_z R_{ij}(x_i W_q(z \mid i), x_j W_q(z \mid j), b - c_q) \Bigg\}. \end{aligned} \tag{5}$$

In Equation 5, the first term stops and selects a terminal action; the second issues another query, with the query minimum interpreted as $+\infty$ when no query is affordable. Thus $R_{ij}(p_i, p_j, C)$ is the minimum regret attainable on the isolated two-hypothesis problem. Replacing d_{ij} by this exact pairwise value cannot weaken the certificate, but its state space may grow exponentially with budget and response branching.

Anytime adaptive bounds. A separate Bellman search operates on the full unnormalized posterior vector. At each unexpanded state, immediate stopping supplies a valid lower bound on adaptive value, while perfect-information utility supplies a valid upper bound. Expanding a state tightens both bounds; complete expansion returns the exact adaptive value. State or time caps therefore produce explicit intervals rather than silently converting an incomplete search into an exact result.

Validation against exact values. Table 5 compares the affinity certificate and its tightenings with exact adaptive values on the 108 synthetic validation cells.

C Noiseless Two-Sided Bracket

For a fixed query set Q, let $\sigma(Q)$ be the number of unordered hypothesis pairs separated by at least one query. A partition of $n = |H|$ hypotheses into m nonempty answer atoms separates at least

$$\sigma_{\min}(m) = \binom{n}{2} - \binom{n-m+1}{2}$$

Table 5: Certificate validation on 108 instances with exact adaptive values; all bounds are sound. Slack is the upper bound minus the exact value (lower is better).

mean slack			fraction of cells		
perfect-info	affinity	exact-pair	exact-pair improves	exact (pair)	exact (best)
0.223	0.166	**0.135**	73.1%	21.3%	**37.0%**

pairs. If $\sigma^\star$ is the maximum pair coverage of any k fixed queries, then

$$N(F,k) \le \max\{m : \sigma_{\min}(m) \le \sigma^\star\}.$$

Pair coverage is monotone and submodular. If greedy selection covers σ_{gr} pairs, its approximation guarantee (Nemhauser et al., 1978) gives

$$\sigma^\star \le \min\left\{\binom{n}{2}, \frac{\sigma_{\mathrm{gr}}}{1-1/e}\right\}.$$

The answer partition produced by the greedy query set supplies a constructive lower bound N_{lb}, while the preceding inequalities supply an upper bound N_{ub}.

For the adaptive side, let S be the surviving hypothesis set and d the remaining depth. Full Bellman expansion obeys

$$U(\varnothing,d)=0, \qquad U(S,0)=1 \text{ for } S\neq\varnothing, \qquad U(S,d)=\min\left\{|S|, \max_{q\in F}\sum_z U(S_{q,z},d-1)\right\}, \tag{6}$$

where $S_{q,z}$ is the answer cell induced by query q. Equation 6 is exact under full expansion. In a truncated search, an unexpanded frontier state is upper-bounded by

$$\min\{|S|, b_{\max}^d\},$$

with value zero when $S=\varnothing$, where $b_{\max}$ is the largest response-alphabet size. Explicitly constructed decision trees provide a lower bound A_{lb}, and the truncated recursion provides an upper bound A_{ub}. Therefore

$$\frac{A_{\mathrm{lb}}-N_{\mathrm{ub}}}{|H|} \le G_{\mathrm{info}} \le \frac{A_{\mathrm{ub}}-N_{\mathrm{lb}}}{|H|}.$$

For any fixed expansion depth, this procedure is polynomial in the explicit signature matrix; complete expansion may be exponential.

D Classical Special Cases

Proof of Proposition 2. Under a uniform prior and exact-identification utility, every nonempty answer atom of a fixed query design contributes exactly $1/|H|$ to Bayes accuracy: the optimal terminal action correctly names one hypothesis in that atom. Likewise, every nonempty terminal cell of an adaptive decision tree contributes exactly $1/|H|$. Maximizing the respective numbers of atoms and terminal cells gives the two value identities and hence their difference.

The following classical combinatorial-search examples explain why richer tools do not necessarily produce greater adaptive value (Katona, 1973; Gerbner & Vizer, 2020; Damaschke, 2019). We use them for interpretation and claim no novelty for these identities.

Proposition 6 (Laminar and unrestricted query families)**.** *Suppose binary queries are represented by subsets of H. If the query family F is laminar, any k pre-committed queries induce at most $k+1$ nonempty answer atoms, so*

$$N(F,k) \le k+1.$$

If F contains a rooted binary hierarchy whose leaves are the singleton hypotheses and whose maximum root-to-leaf depth is at most $\lceil \log_2 n \rceil$, an adaptive policy can nevertheless identify $n=|H|$ hypotheses within that budget. At the other extreme, if $F=2^H$, then pre-committed queries can assign every hypothesis a binary code, giving

$$N(F, \lceil \log_2 n \rceil) = n$$

and therefore zero adaptive information value.

Proof. Selected sets from a laminar family form an inclusion forest. Every nonempty answer atom is either the portion of one selected set not covered by its selected children or the complement of all maximal selected sets, yielding at most $k + 1$ atoms. The stated hierarchy follows one child after each response and reaches a singleton within its maximum depth. For the unrestricted family, choose one set for each bit of a hypothesis index; the pre-committed response vector is then a unique binary code.

Proof of Proposition 5. Index H by $\{0, \dots, n-1\}$ with $n = 2^k$. Dyadic intervals are pairwise nested or disjoint, so the family is laminar and, by the preceding proposition, any k pre-committed queries induce at most $k + 1$ nonempty atoms; by Proposition 2, $V_{\mathrm{na}}(k) \le (k+1)/n$. The dyadic intervals also form a rooted binary hierarchy of depth k with singleton leaves, so adaptive bisection identifies every hypothesis and $V_{\mathrm{ad}}(k) = 1$. For $F = 2^H$, the binary code above gives $N(F, k) = n$, hence $V_{\mathrm{na}}(k) = 1$ and $G_{\mathrm{info}}(k) = 0$.

E Controlled-Experiment Details

Topology construction and latent spaces. Internet Topology Zoo graphs (Knight et al., 2011) are reduced to their largest connected component, stripped of self-loops, and relabeled with consecutive integer node identifiers. The primary exact-identification space contains single-bridge failures: removing a non-bridge does not change reachability and is therefore indistinguishable from no failure under the primary diagnostic interfaces. Headline real-topology panels require at least 12 identifiable failures and use explicit node and query-menu caps when needed for exact search. The constructed panel uses 20-node cores from Erdős–Rényi, Barabási–Albert, Watts–Strogatz, random-geometric, and stochastic-block families, each augmented with the same 12 seeded pendant faults.

Exact optimization. For deterministic responses, each query subset maps every hypothesis to an answer signature. The number of distinct signatures gives the fixed-policy capacity $N(F, k)$. Adaptive capacity $A(F, k)$ is computed by memoized Bellman recursion on the surviving hypothesis set and remaining budget. A value is reported as exact only when both fixed and adaptive searches certify completion. Menu caps, state caps, greedy designs, and truncated adaptive trees remain explicitly labeled as lower or upper bounds.

F Robustness Beyond the Primary Setting

Larger failure spaces. On a separate predeclared set of 40 real topologies, the neighbor-readout gap remains exactly zero at budgets two and three both for all two-link failures and for the unrestricted null-or-single-link failure space. All 160 topology–hypothesis-space–budget cells complete exactly. This extends the structural-zero result beyond the primary single-bridge setting, but not to arbitrary utilities, response channels, or correlated failure processes.

Uncapped query menus. The noiseless bracket certifies zero adaptive information value on every tested network at budgets two and three for uncapped structural menus averaging 26.5 queries, and on 83% and 80% of networks for full menus combining three tool families and averaging 79.5 queries.

Exact optimization with noisy responses. For a uniform prior and exact identification, define the scaled fixed-policy capacity

$$N_W(F, k) = \max_{q_1, \dots, q_k} \sum_{z_1, \dots, z_k} \max_{h \in H} \prod_{t=1}^{k} W_{q_t}(z_t \mid h).$$

The corresponding fixed accuracy is $N_W(F, k)/|H|$. Although this expression uses exactly k queries, it equals the at-most-k optimum because a terminal decision rule may ignore any additional responses. For an unnormalized belief vector w, adaptive capacity satisfies

$$\mathcal{A}(w, 0) = \max_h w_h, \qquad \mathcal{A}(w, d) = \max \left\{ \max_h w_h,\ \max_{q \in F} \sum_z \mathcal{A}(w \odot W_q(z \mid \cdot), d-1) \right\}.$$

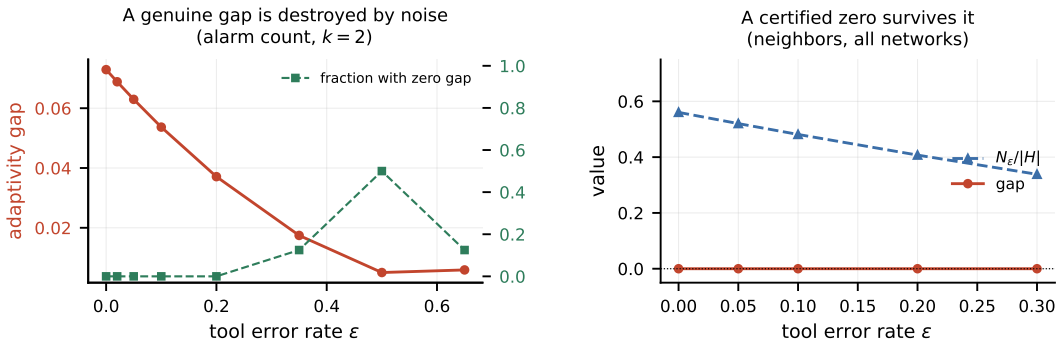

Figure 3: Adaptive information value under noisy response channels. **Left:** on eight alarm-count networks, the mean gap decreases as symmetric response error increases, and half the cells become exact zeros at error 0.50. **Right:** on 25 single-bridge networks, the neighbor-readout gap remains zero through error 0.30 even as identification accuracy falls substantially.

Table 6: Scope of the principal empirical claims. Exactness refers to the corresponding Bayes quantity, not to external validity.

study	readout and units	budget	reported quantity	status
constructed five-family	neighbor; 100 graphs	3	mean gap 0.0000	exact
constructed five-family	alarm count; 100 graphs	3	mean gap 0.1607	exact
real-topology multifault	neighbor; 40 graphs	2, 3	gap 0.0000	exact
noisy unrestricted failures	neighbor; 20 graphs	2, 3	mean gap 0.0024–0.0154	exact
substantive certificate panel	two readouts; 72 cells	2–4	certificate verdicts	bound
Adaptive Elicitation	four-symbol responses; 24 panels	0–4	gap 0.0000–0.0479	exact
GRASP capped-radius counterfactual	edge-ball; 24 strata	2, 3	gap 0.0000	exact
model confirmation	neighbor; 63 lineages	3	zero oracle gap; estimated model effect	mixed
model confirmation	alarm count; 63 lineages	3	positive oracle gap; estimated model effect	mixed
multi-interface follow-up	five readouts; 15 lineages	3	slope 0.2082	estimate
absolute-accuracy panel	two readouts; 25 graphs	2, 3	fixed and adaptive accuracy	exact

Starting from $w = \mathbf{1}$, adaptive accuracy is $\mathcal{A}(\mathbf{1}, k)/|H|$. These recursions reduce to deterministic atom counting when the response noise is zero.

Under symmetric response noise (Figure 3), the mean alarm-count gap on eight real topologies falls from 0.0729 at zero error to 0.0050 at error 0.50, while half of the cells become exact zeros. On 25 single-bridge topologies, the neighbor-readout gap remains zero through error 0.30 while mean accuracy falls from 0.56 to 0.33. However, zero is not channel invariant. A 120-cell exact audit over 20 topologies, two budgets, and three unreliable channels uses 4% false alarms with 15% misses, 10% erasure, and their combination. The resulting mean gaps range from 0.00240 to 0.01538, with 12–19 of 20 topology cells positive, depending on channel and budget. Thus robustness claims must name the response channel rather than only the interface.

G Claim-Scope Crosswalk

Repeated interface names do not identify the same estimand unless the dataset, hypothesis space, prior, utility, response channel, query menu, and budget also agree. Table 6 records the principal scopes. "Mixed" indicates that exact oracle quantities and estimated model effects appear in the same study.

Table 7 summarizes the conclusion each study supports, separating computational exactness from external validity and identifying which study contains model outcomes.

Table 7: Evidence scopes. "Exact" refers to Bayes-value computation, not external validity. Only the final row contains language-model outcomes.

study	scale	supported conclusion
certificate validation	108 cells	implementation accuracy; certificate tightness
controlled interface	100 graphs	the response alone creates or removes adaptive value
real-topology design	25 graphs	zero gap is not saturation; aggregation is not truncation
noise robustness	3 panels	adaptive value depends on the full response channel
Adaptive Elicitation	24 panels	non-graph source-faithful oracle counterfactual
native-source audits	artifacts	provenance and distribution scope; no model outcomes
model confirmation	63 lineages	selected models use externally optimized adaptive evidence

Table 8: Exact Adaptive Elicitation response-matrix counterfactual. Intervals are panel-bootstrap intervals conditional on the pinned matrix. This is not an evaluation of the released learned policy.

follow-up questions	mean adaptive gap	95% interval
0	0.0000	exactly zero
1	0.0000	numerically zero
2	0.0102	[0.0080, 0.0123]
3	0.0295	[0.0255, 0.0339]
4	0.0479	[0.0419, 0.0543]

H Source-Faithful and Native Audits

We distinguish three provenance levels. A *native-source audit* executes released code or directly parses released data and schemas. A *source-faithful counterfactual* preserves a released response matrix or accessor but supplies an author-defined prior, utility, latent space, or query cap. A *proxy* preserves only a documented tool family on a shared task. These levels support different conclusions and are not pooled as native benchmark performance.

H.1 Adaptive Elicitation

From the pinned TwentyQuestions release of Adaptive Elicitation (Wang et al., 2025), we parse a test matrix containing 170 entities, 878 question columns, and four response symbols after removing the identity field. Each of 24 disjoint panels contains five target questions and 21 candidate questions. Both policies receive one random initial answer; the pre-committed policy fixes all follow-up questions in advance, whereas the adaptive policy conditions each choice on previous responses. The prior is empirical-uniform over entities, and utility is mean exact accuracy on the five target questions. Table 8 reports the complete follow-up-budget progression.

H.2 GRASP

We execute the released GRASP graph-information tool rather than reimplementing its behavior (Guo et al., 2026). The tool returns all edges in the chosen node's radius-r neighborhood, while the prompt discloses the graph size and permits an unrestricted integer radius. A radius-$(N-1)$ query therefore recovers every connected N-node graph in one pre-committed call. This argument does not extend to disconnected episodes. Table 9 quantifies this connectivity boundary.

Table 9: GRASP native connectivity audit. A radius-$(N-1)$ query recovers the whole graph only in connected episodes.

released distribution	graphs	connected	maximum components
builder-eligible public test set	4,700	2,866 (60.98%)	10
builder's fixed blind-test sample	470	286 (60.85%)	8

On disconnected graphs, observing one component can inform an adaptive policy which disclosed node remains unvisited, while pre-committed centers may fall in the same unknown component. We therefore do not make a benchmark-wide one-call zero-gap claim. In a separate source-faithful counterfactual that caps the native edge-ball radius at one or two hops over 24 released-data strata, all 96 stratum–radius–budget cells at budgets two and three have exact zero gap. This counterfactual preserves the released accessor but uses our prior and graph-global utility.

H.3 GDS-AGENT

The GDS-Agent audit examines the released tool schema and result-limit implementation (Shi & Panagiotas, 2025); full execution would require a live Neo4j Graph Data Science deployment. Its relationship-stream operation returns source, target, and relationship type for the in-memory graph. The shipped defaults allow 500 rows, 100,000 total characters, and 200 characters per cell.

The released London graph contains 302 stations, 406 connections, and 13 lines. One unfiltered relationship-stream call returns all 406 rows in 3,858 characters, so none of the shipped limits binds. Under our uniform single-connection-failure counterfactual, this one call reveals the complete graph, giving

$$V_{\mathrm{na}}(C) = V_{\mathrm{ad}}(C) = 1$$

for budgets one through three. This is a saturation zero under an author-defined latent task and utility, not a native model-performance result.

I MODEL-STUDY DETAILS

Arms and evidence generation. Each episode provides the graph adjacency list, candidate failures, and diagnostic interface contract, but no symptom that directly reveals the true failure. In the primary arms, an exact external optimizer generates either an optimal adaptive transcript or an optimal pre-committed transcript. Both are then presented in the same final-prediction format. Additional arms present identical fixed evidence jointly or sequentially, allow models to select their own queries, or supply the posterior candidate set. These arms separately probe evidence quality, presentation, query-policy discovery, and final decision competence.

Serving and scoring. Models are served locally with vLLM (Kwon et al., 2023) from AWQ-quantized checkpoints. Final decoding is constrained to a single answer index. Intention-to-treat scoring accepts exactly one in-range candidate index; extra text, missing or out-of-range indices, truncation, and forced-completion retries are scored as incorrect. This rule is fixed across models, interfaces, and arms. In the four primary arm-by-interface cells, mean input-plus-output token counts are 3,820 and 3,806 for pre-committed and adaptive alarm-count evidence, and 4,058 and 3,963 for the corresponding neighbor-readout arms; configured output opportunity differs by at most one token. These checks do not make the prompts computationally identical.

Sampling and inference. The 63 evaluation lineages are disjoint from 18 calibration lineages and all pilot lineages. Every model–interface–seed block contains 12 unique truth draws per lineage, giving 24 paired repeats across two seeds. Outcomes are averaged within each model and lineage and then across the frozen three-family roster, leaving the lineage as the independent unit. The unstudentized test statistic is the mean roster-averaged lineage effect. Confirmatory inference uses 100,000 independent lineage-level sign reversals and percentile confidence intervals from resampling lineages with replacement. The equivalence endpoint uses two one-sided lineage-level tests at the prospectively fixed margin; its displayed 95% interval is descriptive rather than the TOST decision rule. Holm correction is applied to the three confirmatory p-values. Prospective timing is self-attested because the repository entered version control after these computations. The oracle gaps are fixed design quantities and are not part of the Holm family. As a post hoc check on the sign-symmetry assumption behind the frozen unstudentized test, studentized one-sample lineage tests give $p = 1.54 \times 10^{-6}$ for the primary effect and $p = 0.00126$ for the interaction. These sensitivity tests do not replace the prespecified analysis.

Heterogeneity and additional controls. The primary alarm-count effect is 0.0324 for Qwen2.5-32B, 0.0423 for Mistral-Small-24B, and 0.1005 for Llama-3.3-70B; leaving out each family in the

Table 10: Fixed-cap parsing audit on a certified-zero interface. The same QwQ-32B episodes and per-call cap are used; changing total-budget accounting changes which arm is resource-starved. Neither row estimates adaptive information value.

accounting	fixed accuracy	adaptive accuracy	difference	95% interval
equal per-call	0.025	0.450	+0.425	[+0.250,+0.550]
equal total	0.525	0.475	-0.050	[-0.225,+0.075]
exact information ceilings	0.784	0.784	0.000	exactly zero

same order gives positive estimates of 0.0714, 0.0665, and 0.0374. The posterior-candidate-set condition reaches 0.520 roster accuracy, showing that the models can use a compact summary of the available evidence.

Qualification. Qualification uses calibration lineages only and requires exact parsing of at least 0.98 and truncation, invalid-query, and runtime-failure rates of at most 0.01 in every arm. Qwen2.5-32B, Mistral-Small-24B, and Llama-3.3-70B satisfy these criteria and form the frozen roster.

J Protocol Deviations and Inconclusive Follow-Up

Qualification deviation. Gemma-3-27B initially failed the truncation requirement. Its required output-cap escalation was mistakenly completed only after confirmatory outcomes had been inspected. The model continued to fail the qualification criteria in every arm at both 1,600 and 3,200 output tokens, so the frozen roster and all reported confirmatory results were unchanged. We nevertheless record this as a temporal protocol deviation rather than describing qualification as perfectly preregistered.

Multi-interface dose response. A prospectively specified follow-up crosses five exact-gap interfaces over 16 frozen topology snapshots. Independent review later established that two Deutsche Telekom snapshots belong to one operator lineage; averaging them leaves 15 independent lineages. Only four of the 80 topology–interface points have positive exact gaps, making the intended gradient weak despite outcome-blind selection.

The clustered slope of model benefit on exact adaptive information value is 0.2082, with 95% bootstrap interval $[-0.0664, 1.1877]$. Spearman correlation is 0.201 with interval $[-0.092, 0.389]$, and one of the three leave-one-model-out slopes is negative. The frozen bootstrap encountered singular resamples after outcomes existed; the repaired procedure deterministically rejected 1,010 of 101,010 attempts to obtain 100,000 valid draws. A post hoc delete-one-lineage jackknife is also inconclusive, with interval $[-0.692, 1.108]$. The follow-up therefore does not establish that larger certified gaps monotonically produce larger model benefits.

K Exploratory Resource-Accounting Audits

Early model experiments used small samples and fixed per-call output caps. We retain them for traceability because they demonstrate how resource accounting can reverse an apparent adaptive advantage, not because they provide confirmatory model evidence. Table 10 shows the resulting reversal, and Appendix M lists the early model results.

Under equal per-call accounting, the fixed arm parses on only 5% of episodes and is truncated on 97.5%. Matching total resources raises its parse rate to 52.5% but instead starves the adaptive arm, whose truncation reaches 95%. The protocol choice moves the observed contrast by 0.475 while the exact information gap remains zero. These failures motivated the qualification gates, constrained final decoding, and matched-resource primary design.

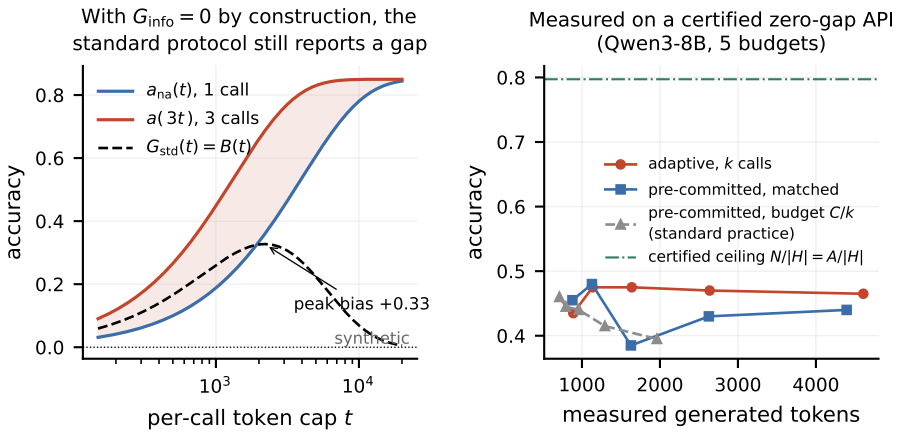

Figure 4: Compute confounding. **Left** (synthetic): accuracy depends only on generated tokens, so $G_{\text{info}} = 0$, yet equal per-call evaluation reports a gain of up to $+0.327$. **Right** (measured, Qwen3-8B, certified-zero interface): adaptive, matched pre-committed, and equal-per-call pre-committed conditions against measured tokens; deviations are optimization or compute effects, not adaptive information value.

Table 11: Qwen3-8B calibration on a certified-zero neighbor readout. Every cell has $G_{\text{info}} = 0$, exact-format parse rate 1.000, and zero truncation. Pointwise p-values are unadjusted; none survives correction over the five planned budgets.

B_{tok}	matched effect	p	equal-per-call effect	p
250	-0.020	0.636	-0.025	0.522
500	-0.005	1.000	+0.030	0.362
1000	**+0.090**	**0.013**	+0.035	0.265
2000	+0.040	0.332	+0.055	0.126
4000	+0.025	0.551	+0.070	0.092

L Compute Sweeps and Zero-Information Calibration

The sweep allocates total generation budget B_{tok} to the adaptive arm, one call of size B_{tok} to the matched pre-committed arm, and one call of size B_{tok}/k to the standard equal-per-call arm. Final-answer allowance is identical. A cell is reportable only when both compared arms parse at least 0.95 and their truncation rates differ by at most 0.10. Across five budgets, two interfaces, and five models, 30 instruction-model cells pass these gates; all 20 reasoning-model cells fail. Table 11 reports the certified-zero calibration curve; Appendix K shows how fixed per-call caps can reverse an apparent adaptive advantage.

Figure 4 illustrates the confound. In the synthetic left panel, accuracy is set to $a(t) = 0.85\,(1 - e^{-t/4000})$ with $k = 3$ calls, so three calls simply receive three times the token allowance. The measured right panel uses Qwen3-8B on Internet Topology Zoo networks (Knight et al., 2011) with the certified-zero neighbor readout, $k = 3$ calls, 200 episodes per point, total budgets of 250–4000 tokens, a 600-token answer allowance, and temperature 0.6. The adaptive condition is below the matched pre-committed condition at the two smallest budgets and above it at the three larger ones.

One budget has a nominal positive matched effect, but it does not survive correction over the five planned budgets. Because the exact information gap is zero in every row, any observed contrast is necessarily attributable to bounded-model optimization, compute allocation, or sampling variation.

Table 12: Model-selected policies on the positive-information readout. Effects compare adaptive with matched pre-committed evaluation.

model	B_{tok}	adaptive accuracy	fixed accuracy	effect	pointwise p
Qwen3-8B	250	0.085	0.145	-0.060	0.023
	500	0.135	0.190	-0.055	0.061
	1000	0.100	0.195	-0.095	0.003
	2000	0.120	0.200	-0.080	0.009
	4000	0.130	0.125	+0.005	1.000
Qwen2.5-14B	250	0.100	0.125	-0.025	0.424
	500	0.095	0.125	-0.030	0.307
	1000	0.100	0.125	-0.025	0.424
	2000	0.095	0.125	-0.030	0.307
	4000	0.095	0.125	-0.030	0.307
Qwen2.5-32B	250	0.180	0.185	-0.005	1.000
	500	0.180	0.185	-0.005	1.000
	1000	0.180	0.185	-0.005	1.000
	2000	0.180	0.185	-0.005	1.000
	4000	0.180	0.185	-0.005	1.000

L.1 Positive-Information Model-Selected Sweeps

Every row below uses the alarm-count readout with exact $G_{\text{info}} = 0.1594$, fixed ceiling 0.384, adaptive ceiling 0.543, exact-format parse rate 1.000, and zero truncation. Budget rows reuse the same episodes and are therefore curve points rather than independent replications (Table 12).

A positive information opportunity does not guarantee successful model-selected adaptive querying. The repeated negative effects, together with the positive externally optimized evidence result, localize the shortfall to policy discovery rather than evidence availability.

L.2 Reasoning-Model Budget Floor

QwQ-32B and DeepSeek-R1-Distill-14B exhaust every tested output budget from 250 to 4,000 tokens and fail the reportability gate in all 20 model–interface–budget cells. QwQ produces no scored answer in any arm. The adaptive parse rate of DeepSeek-R1-Distill rises only from zero to 0.07 across the tested budgets. These are unmeasurable protocol cells rather than evidence of zero task accuracy or zero adaptive capability. Their failure motivated excluding the original reasoning-model sweeps from the core model claim.

M Early Exploratory Model Results

Early experiments used 40–60 episodes and, for instruction-tuned models, 320–400 output tokens. QwQ-32B used 40 episodes and a 4,000-token cap under matched accounting. These cap-sensitive comparisons are retained for traceability and are superseded by the independent-lineage confirmation (Table 13).

No model shows a reliable positive model-selected adaptive effect on the positive-gap readout. Supplied evidence can improve performance, but the pattern is not monotone in model size. These exploratory observations motivated the larger externally optimized evidence study.

N Power Planning

For E paired binary episodes with discordant-pair probability π, the usual null-variance normal approximation for matched binary outcomes (Miettinen, 1968; Connor, 1987) gives

$$\text{MDE}(E) = \left(z_{1-\alpha/2} + z_{1-\beta_{\text{II}}}\right)\sqrt{\frac{\pi}{E}}.$$

Table 13: Exploratory model results at query budget three. "Supplied evidence" uses externally chosen structural queries. These results are not confirmatory.

model	readout	fixed	adaptive	effect	supplied evidence
Qwen2.5-1.5B	neighbor	0.067	0.067	+0.000	0.050
	alarm count	0.067	0.050	-0.017	0.067
Qwen2.5-7B	neighbor	0.017	0.083	+0.067	0.017
	alarm count	0.133	0.100	-0.033	0.217
Qwen2.5-14B	neighbor	0.183	0.133	-0.050	0.167
	alarm count	0.133	0.067	-0.067	0.150
Qwen2.5-32B	neighbor	0.267	0.283	+0.017	0.317
	alarm count	0.183	0.133	-0.050	0.217
QwQ-32B	neighbor	0.525	0.475	-0.050	0.675
	alarm count	0.225	0.150	-0.075	0.275

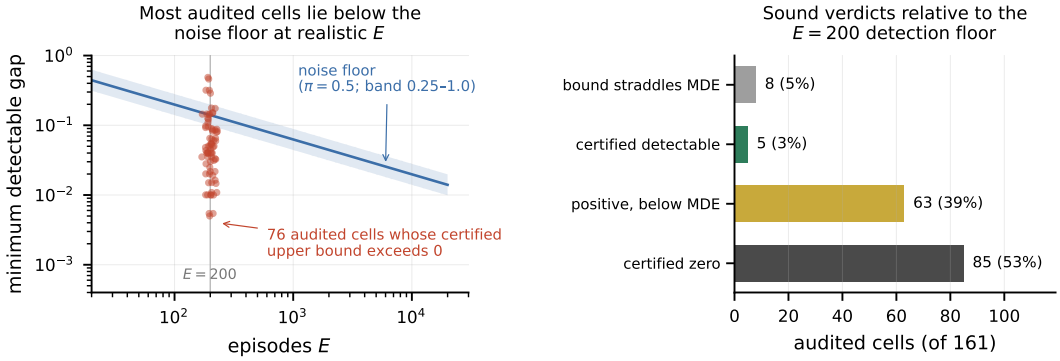

Figure 5: Power-planning sensitivity. **Left:** approximate detectable effect versus evaluation size under alternative discordance assumptions. **Right:** audited cells grouped by sound verdict relative to the illustrative minimum detectable effect (MDE) at $E = 200$. "Straddles" means that the certified interval crosses this threshold. Categories depend on the planning assumptions.

This follows by writing the paired difference as $(b - c)/E$ and conditioning on $b + c$, whose null variance is $E\pi$. It is a planning approximation rather than an exact detectability theorem because π is unknown before evaluation.

At 80% power, two-sided $\alpha = 0.05$, and $\pi = 0.5$, detecting an effect of 0.10 requires approximately 390 paired episodes, while detecting 0.02 requires approximately 9,800. Among 161 audited cells with query budget at least two, 85 have certified upper gap zero, 63 have positive gaps whose upper bounds remain below the $E = 200$ planning threshold, five lie above it, and eight straddle it. These categories are sensitivity calculations, not impossibility claims. Figure 5 visualizes this dependence on the planning assumptions.

O Correction and Traceability Ledger

We retain corrected and failed analyses because they delimit what the evidence supports.

- An earlier ridge-table gap of $+0.036$ arose when an exact fixed-design request silently fell back to a greedy lower bound. Exact enumeration gives $N/|H| = 1$ and gap zero. Exact requests now fail explicitly when search does not complete.
- A historical quantity labeled as "$p = 1.000$" was a bootstrap direction frequency rather than a p-value. Exploratory episode contrasts were recomputed with exact McNemar tests; confirmation uses lineage-level inference.
- A binary 2^k ceiling was incorrectly applied to multivalued tools. Current bounds use observed response alphabets or explicit Bellman frontiers.

- Three auxiliary scripts used process-dependent hash values to encode answer sets. Equality-based interning removed this nondeterminism and left all reported capacities unchanged.
- Equality of adaptive and pre-committed optima was previously interpreted as requiring equal bounded-model protocol accuracy. Equation 2 gives the corrected interpretation.
- Exact historical byte snapshots of the confirmatory generation runner were not recovered. Raw outputs can be rescored and the complete lineage analysis rerun, but model generation is not claimed to be byte-for-byte reproducible.