LEARNING DEEP CHOICE HIERARCHIES FROM FOUR EXACT PROBABILITY RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Can four probability responses reveal an unknown deep choice hierarchy, and can an inexact reconstruction be certified? We study reduced nested-logit trees of arbitrary finite arity with unknown parameters and affine item-input scales, under independently adjustable continuous inputs and full positive probability-vector access. A scale-free finite-coupling identity reuses the same responses across all levels. A baseline and three random perturbations identify the entire model almost surely; for three or more alternatives, four are optimal even with adaptive queries. For inexact observations, a candidate-only shared-response certificate uses the proposed tree's clades to exclude every competing compatible topology, with cubic arithmetic work after a strict binary proposal is supplied. When unique structure cannot be certified, a separate packed outer cover can instead certify one common predictor across structures and all available-item sets on a declared domain. These guarantees require family membership, valid observation intervals, and successful verified computation; refusal is permitted. Across 12 fresh synthetic models and 72 correlated design/precision conditions, shared-response certification provides 66 total-variation bounds at most 0.01, versus 36 for packed certification. A matched enclosure ablation supports the mechanism. Public-component polytomies require the complementary packed route. Float32 certification remains weak, a dense classical control is faster locally, and frequency-based structure recovery fails the prespecified criterion under the tested sampling conditions. This is a probability-query boundary with conditional certification, not recovery from four choices or a general noisy-sample guarantee.

1 INTRODUCTION

How many probability queries reveal an unknown hierarchical choice model when the input units are unknown? This is an identification question about a promised-family interface, not learning from four observed choices. A response contains a full positive n-dimensional probability vector; the learner can independently perturb each item's continuous input. Such access separates a simulator or prediction interface's information from the cost of estimating its outputs. It is stronger than behavioral data access, but does not reveal the tree, gradients, or input calibration.

We identify a reusable structure: a two-coefficient finite-coupling identity whose output is itself a valid input message at the next level. Unknown parent temperatures never have to be guessed in this bottom-up calculation. The same baseline and three perturbations therefore suffice at every depth, with unknown affine input scales and arbitrary finite branching. A matching lower bound, even for adaptive queries at every $n \geq 3$, uses overlapping conditional response charts, not merely parameter counting. For inexact outputs, we turn a proposed strict tree into a checkable certificate excluding every competing compatible structure, without enumerating all subsets. When that route refuses, a complementary packed cover can certify a common predictor even when compatible structures differ.

The finite-coupling perspective is motivated by the power-law and finite-variation formulation of relative N-level operators (Yu, 2025). Our contributions are the unknown-scale contraction, its sharp probability-query boundary, and two complementary conditional certificates: candidate-only structural exclusion and common prediction under ambiguity. Nested logit, odds-ratio cancellation, log-sum-exp, analytic zero-set arguments, interval enclosures, and nested entropy are established tools,

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.

not new operators claimed here. We preserve the counterevidence: classical queries can be faster overall, low-precision certificates can be useless, and the tested frequency-learning extension fails.

2 INFORMATION MODEL AND CLOSEST WORK

A rooted tree has labeled leaves $[n]$, internal arity at least two, and inverse temperatures $k_v > 0$, with $\tau_v = 1/k_v$. For independently adjustable inputs $x \in \mathbb{R}^n$, define

$$f_i = b_i + a_i x_i, \quad a_i \neq 0, \qquad f_v = k_v^{-1} \log \sum_{c:\mathrm{pa}(c)=v} e^{k_v f_c}. \tag{1}$$

The oracle returns leaf probabilities given by products of the conditional softmax probabilities along each root-to-leaf path. It supplies no derivatives, latent utilities, or parameters. Responses are exact for identification; the separate numerical result assumes valid enclosing intervals. Queries lie in an open product domain. A *reduced* tree has unequal temperatures on adjacent internal nodes; equal-temperature edges collapse without changing its probability function.

Common utility scaling and shifting are symmetries. Set $k_{\mathrm{root}} = 1$ and a bias gauge such as $b_n = 0$ for exact full-model recovery. The interval algorithm equivalently sets the baseline root value to zero. These are normalizations, not extra observations. The upper bound allows unordered temperatures. For the lower bound and certificates we use the random-utility-compatible subclass $0 < \tau_{\mathrm{child}} \leq \tau_{\mathrm{parent}}$ on internal edges, with strict order where stated. Full-model recovery includes the tree and canonical parameters, hence the entire probability function, not only a nest partition.

What differs from existing structure learning? Classical nested logit already has within-nest odds identities (Train, 2009; Kovach & Tserenjigmid, 2022); Kovach & Tserenjigmid (2022) also characterize multistep endogenous structures. Mixed-integer likelihood model selection uses a different data model (Aboutaleb et al., 2020). IIA-oriented learning (Benson et al., 2016) and all-slate distribution approximation from conditional samples (Chierichetti et al., 2026) use other primitives. Most directly, Yu et al. (2026) use experimental assortments, probability ratios, recursive nest identification, parameter recovery, and finite-sample analysis, including a real deployment and a recursive deep-tree discussion in their Appendix C. We do not claim that recursion or probability ratios are new, or compare four continuous-input queries to their assortment count as if the oracles were equal. Randomized price experiments address another statistical regime (Li et al., 2015). Prediction-interface extraction also predates this work (Tramèr et al., 2016). Our boundary concerns unknown affine scales and independent continuous inputs, not improved customer-sample complexity. Nested-entropy learning results (Martin et al., 2022; Mertikopoulos & Sandholm, 2024) supply the classical variational structure used below.

Two information interfaces, not a query-count race. The following comparison makes the access assumptions explicit. The assortment result is its analyzed nest-identification setting; its recursive deep-tree discussion is not the same theorem as ours.

	This paper	Yu et al. (2026)
Intervention	Independent continuous item inputs; all items present during learning.	Availability sets change; underlying preference parameters stay fixed.
Observation	Full exact probability vector, or valid enclosing intervals for certification.	Assortment choice probabilities; finite-sample estimation is also studied.
Unknowns	Hierarchy, temperatures, biases, and nonzero affine input scales.	Nests and preference parameters; not calibration of our continuous-input coordinates.
Exact boundary	Four responses for full canonical recovery, even adaptively, at every $n \geq 3$.	$O(\log n)$ nest-identification assortments and an $\Omega(\log n)$ adaptive lower bound under their assumptions.

Neither result dominates the other without an oracle reduction. Our same-access and identical-transcript controls in Section 5 address comparisons within our interface instead.

3 SCALE-FREE CONTRACTION AND A SHARP EXACT BOUNDARY

Write $\Delta h = h(x) - h(x^0)$ and $P_A = \sum_{i \in A} p_i$ for a true current group. At a leaf set $u_i = \Delta x_i$; an internal group has message $u_A = k_A \Delta f_A$. A binary contraction forms $C = A \cup B$ from two groups sharing a current nest. In a binary tree C is their true parent; in a multiway nest it can be a partial aggregate with that nest's temperature. Then

$$\Delta \log(P_A/P_B) = c_A u_A - c_B u_B, \qquad c_A = \begin{cases} k_C a_i, & A = \{i\}, \\ k_C/k_A, & A \text{ internal.} \end{cases} \tag{2}$$

Define c_B analogously. With $P_C = P_A + P_B$, $w_A = P_A(x^0)/P_C(x^0)$ and $w_B = 1 - w_A$, its aggregate message is

$$u_C = \log(w_A e^{c_A u_A} + w_B e^{c_B u_B}). \tag{3}$$

Two coefficients can be fitted on two perturbations and tested on a third. Crucially, the result closes under composition: it is $k_C \Delta f_C$ without knowing k_C. The same three changes enter every contraction.

A three-item closure example. Consider the hierarchy $((1,2),3)$ at $x^0 = 0$ with baseline probabilities $(0.2, 0.3, 0.5)$, $\tau_{12} = 1/2$, $\tau_{\text{root}} = 1$, and slopes $(1, 3/2, 4/5)$. These are the example's hidden parameters, not learner inputs. At the first merge, the learner obtains coefficients $(2,3)$ from changes of $\log(p_1/p_2)$. The baseline weights $(0.4, 0.6)$ then give the reusable message

$$u_{12} = \log\big(0.4e^{2\Delta x_1} + 0.6e^{3\Delta x_2}\big), \qquad \Delta \log \frac{p_1 + p_2}{p_3} = \tfrac{1}{2} u_{12} - \tfrac{4}{5} \Delta x_3.$$

Thus the next merge learns $(1/2, 4/5)$ without having known τ_{12}. Normalizing the root temperature to one recovers $\tau_{12} = 1/2$ and slopes $(1, 3/2, 4/5)$; baseline conditional weights recover the biases up to the common shift. For example, the three perturbations $0.1e_1, 0.1e_2, 0.1e_3$ give rank two for both true-pair designs, while each false leaf pair fails the three-equation consistency test. This is an illustrative design for this target, not a universal deterministic design. The structural step is that u_{12} replaces two unknown utilities in the *same* three-change calculation, rather than requiring new responses at the next depth.

Exact algorithm. Query a baseline x^0 and three independent inputs with continuous densities on an open product domain. Start with singleton groups and three-component u_i vectors. Test active pairs for a rank-two consistent solution to the three equations in (2). Contract any passing pair by (3), saving its coefficients. At the root set $k_{\text{root}} = 1$ and recover internal $k_A = k_C/c_A$, leaf $a_i = c_i/k_C$, and $f_A(x^0) = f_C(x^0) + k_C^{-1} \log w_A$. Recover biases and apply the gauge. For multiway nodes, passing pairs can introduce equal-temperature binary refinements; collapse those edges at the end.

Theorem 1 (Exact response boundary)**.** For each fixed reduced model (1), $n \geq 3$, a baseline and three independent continuous perturbations recover its full canonical model almost surely. This holds for arbitrary finite internal arity and nonzero unknown slopes. For each $n \geq 3$, no measurable randomized adaptive algorithm with at most three full probability responses succeeds almost surely on every model, even in the strictly compatible binary subclass with positive slopes. Thus the exact full-model optimum is four even with adaptive queries.

Upper-bound mechanism. A false sibling pair's augmented three-column design has generic rank three. Differentiate its odds ratio in an outside group's input at the least common ancestor. The obstruction uses an *adjacent* internal temperature difference; nonadjacent labels alone can cancel. True-group features depend on disjoint inputs and are independent. Nonzero analytic determinants vanish only on null sets (Mityagin, 2015). A finite union over all true frontiers and compatible contraction histories handles data-dependent merging without assuming fresh independence along the selected path. For multiway nodes, permit unions of whole children and later collapse artificial equal-temperature edges. Appendices B–C give the full argument.

Lower-bound mechanism. Fix a proper full-support prior before all policy coins, then condition successively on the first two responses. Dependent leaf perturbation pairs leave positive-dimensional parameter fibers. Otherwise three binary refinements of a common ternary root have regular inverse charts within the *same fixed-prefix fiber*. At least two strict-RUM sides overlap, so different full

functions yield the same adaptive transcript. A conditional construction also handles an untouched input coordinate; adding known alternatives to a three-item example would not suffice. Integration and Fubini give a fixed hard target, not one selected after the coins (Appendix G). Appendix F retains the simpler nonadaptive argument.

Each response has $n - 1$ independent real coordinates: four is neither a four-bit nor a four-sample claim. The ideal cached contraction uses at most $(n - 1)^2$ pair tests. The floating implementation instead uses residuals and a heap, and is not an exact decision procedure. Passing-pair storage can be quadratic for multiway models. Theorem 1 provides no uniform bit-complexity or conditioning guarantee.

4 From exact recovery to checkable certificates

Assume valid positive output intervals, weakly compatible models, and a bounded input box $\mathcal{X}$. A greedy minimum-residual tree is not a certificate. The exact boundary, candidate-only verification, and packed common prediction have different guarantees:

Route	Certified conclusion	Essential boundary
Exact four responses	Full canonical model, arbitrary finite arity.	Exact vectors and generic queries; no uniform precision.
Candidate-only graph	All compatible models share one strict binary tree and enclosed parameters.	May refuse; does not establish existence or handle polytomies.
Packed cover	One predictor close to all retained compatible models, even across structures.	May be unresolved; worst-case exponential search.

Recovery–verification workflow. Input: four query locations, positive response intervals and a prediction box; family membership and interval validity are assumed, not tested.
1. Generate a strict binary proposal from the interval midpoints; failure is permitted.
2. Verify *all* fixed inverses, coefficient margins and nonsibling determinants on the shared-response box. Only if every check passes, return the proposal with all-compatible-model parameter and prediction bounds.
3. Otherwise report *not certified*, or try the separate packed route for a common predictor. Distinguish an empty cover from an unresolved search. A refused certificate is not proof of nonexistence; successful certification does not establish family membership or existence.

Candidate-only exclusion (Proposition 18). Supply any strict binary candidate with clades $\mathcal{C}_0$. Enclose the observation intervals in one relative-response box about its exact responses, and express every message as a function of these same $4n$ variables. For each candidate sibling, fix two rows of (2), solve for its coefficients, and propagate (3). Verify on the *whole box*: every fixed inverse stays nonsingular; internal-child coefficients stay in $(0, 1)$ and leaf coefficients exclude zero; and every disjoint nonsibling $A, B \in \mathcal{C}_0$ satisfies

$$0 \notin \det[\, u_A, -u_B, \Delta \log(P_A/P_B)\,].$$

If all checks pass, *every* compatible reduced weak-RUM model has the candidate's strict binary topology and enclosed parameters. The proof selects a smallest noncandidate clade in any competing binary refinement: its two child subtrees already coincide with candidate subtrees, so their true sibling relation forces a determinant excluded by the test to be zero. This rules out all competing structures, not merely errors along a greedy path. It does not prove a compatible model exists.

Natural interval ranges are intersected with centered enclosures using verified derivatives over the full response box; the same observation variables persist across depth. Centered forms and interval differentiation are classical (Rump, 2010); the structural exclusion enabled by message closure is the contribution. There are $O(n^2)$ pair tests with $4n$ derivative components, hence $O(n^3)$ arithmetic work *given* a proposal, not bit complexity or a uniform noisy-learning bound. Appendix Q specifies the test and proves it; a failed check means refusal. An explicit ideal midpoint least-squares proposal makes the full proposal–verification procedure locally complete for each fixed strict binary target and a generic four-input design (Corollary 19); the fixed-threshold floating fitter used in experiments has no such universal guarantee.

Common prediction when structure remains ambiguous. For every disjoint subset pair, the verified algorithm encloses a two-coefficient inverse with outward arithmetic, intersects available inverse boxes, and checks every original equation. It removes a branch only on a certified contradiction. A potentially feasible branch without a certified inverse, or a resource cap, yields *unresolved*, even if another branch has already reached the root.

A subset directed acyclic graph stores accepted splits. All child-message images are hulled before larger subsets are tested; a top-down pass hulls parameter images over *all* reachable parent paths. A binary derivation belongs to an outer cover, not necessarily to the feasible-model set. Packing avoids explicit enumeration in favorable cases but is worst-case exponential.

Theorem 2 (Outer cover and common prediction)**.** Conditional on model-family membership, valid intervals, sound arithmetic, and completed search, every compatible model has a represented binary refinement with its canonical parameters enclosed. Construct a reference hierarchy from clades present in every full derivation. Assign other groups their smallest containing common clade's reference temperature. Equal-temperature collapse then gives the same reference function in every derivation. If e bounds its leaf-utility difference from every enclosed model on $\mathcal{X}$, δ bounds corresponding temperature differences, and $t_{\min} > 0$ bounds all compared temperatures below, then

$$\sup_{\substack{\theta \text{ compatible},\ x\in\mathcal{X} \\ \varnothing\neq S\subseteq[n]}} \mathrm{TV}(p_{\theta,S}(x), p_{\mathrm{ref,S}}(x)) \le \min\left\{1, \frac{e+\delta C_n}{2t_{\min}}\right\}, \quad C_n = \sqrt{(\log n)^2 + 2\log n + 2}. \tag{4}$$

This is not a family goodness-of-fit test or a way to establish interval coverage. Available-item sets S prune the same hierarchy. The reference need not fit the observations itself: it is a single function close to all compatible models. Appendix I proves the common-reference construction. An empty cover, an unresolved search, and a completed nonempty cover are distinct outcomes.

Decision meaning of a certificate. For a common feasible decision family $a \mapsto (x(a), S(a))$ within the certified domain and shared rewards $0 \le r_i(a) \le R$, let $V_\theta(a)$ denote expected reward. A uniform TV bound B gives $|V_\theta(a) - V_{\mathrm{ref}}(a)| \le RB$ for every compatible θ. A ξ-optimal reference decision ($\xi \ge 0$) therefore loses at most $2RB+\xi$ against the compatible model's supremal value over that same family (Appendix E). Thus structure ambiguity need not preclude certified model-based decisions; this is not a decision solver, causal effect or real-user reward guarantee.

Avoiding products of layerwise bounds. The classical nested entropy is $\mathcal{H}_\tau(p) = \sum_v \tau_v P_v H((P_c/P_v)_c)$. On every simplex tangent z it satisfies

$$-z^\top \nabla^2 \mathcal{H}_\tau(p) z \ge t_{\min} \sum_i z_i^2/p_i.$$

The nonnegative conditional-entropy curvature terms telescope, and unary nodes left by pruning contribute zero. A temperature perturbation's gradient is bounded by $\delta(-\log p_i)$. The surprisal second moment is at most $(\log n)^2 + 2\log n + 2$. Weighted Cauchy–Schwarz and path integration yield (4). Appendix H proves the bound and gives a fixed-root example showing that logarithmic width dependence cannot generally be removed. A tiny $t_{\min}$ or broad parameter hull can still make it useless. Uncertain slopes also prevent a general all-$\mathbb{R}^n$ result: the input domain matters.

A relative-temperature refinement. When both compared models obey the temperature order, let $\rho = \max_v |\log(\tau_v/\bar{\tau}_v)|$ and $t_{\mathrm{ref}} = \min_v \bar{\tau}_v$. For the same common predictor, Appendix P proves the alternative bound $\min\{1, e/(2t_{\mathrm{ref}}) + \rho \log n\}$. The temperature term is not inflated by an unrelated, exactly known cold nest. The utility term remains conditioned, relative errors can be large near zero, and ordering is essential. This refines the certificate, not identification or noisy-sample recovery.

Proposition 3 (Local finite-precision bridge)**.** Fix a strictly ordered binary model, a bounded prediction box, and a generic four-input design. For complete packed search in exact-real interval arithmetic, there are model-and-design-dependent $\epsilon_0 > 0, C < \infty$ such that valid output widths at most $\epsilon < \epsilon_0$ yield precisely $2n-1$ subset states, the unique true full topology, parameter widths $O(\epsilon)$, and common-reference TV bound at most $C\epsilon$.

This is local consistency, not a uniform noise guarantee or a new interval-analysis principle. Appendix L handles all disjoint true-clade pairs, not just one path. Finitely many nonzero inverse

minors and false-pair residuals persist under small perturbations; smooth interval propagation has linear width locally. Convergent finite arithmetic adds a rounding-enclosure floor. The constants can deteriorate near collapse or singular designs. Polytomies are excluded from the linear-state and unique-binary-topology conclusion. For fixed reduced weak-RUM polytomies, complete packed search still has a local $O(\epsilon)$ common-predictor certificate (Appendix S), but a flat nest can retain $2^n - 1$ subsets. Neither result guarantees completion for ActivitySim under a fixed resource budget.

5 Evidence, refusals, and conditioning

These experiments audit scope, not population performance on behavioral datasets. Learning queries are counted separately from teacher generation and held-out checks. Audit parameters are never inference inputs. Full original tables, seeds, and correction records appear in Appendices K and J.

5.1 Fresh-model certification and a matched mechanism test

A protocol frozen before execution uses 12 new teachers: 8/16 leaves, comb/random-split trees, and three parameter seeds, with two input designs and budgets $b = 20, 24, 40$. The 72 conditions are correlated, not 72 independent models. Both methods receive the same four responses in valid intervals of relative budget 2^{-b} and use the same prediction box and all menus. They retain nominal 20-second cooperative checks with different timed scopes, not matched end-to-end budgets (Appendix R). These budgets are not claims about actual float32 outputs.

Table 1: Fresh 72-condition core. Useful means conditional TV bound ≤ 0.01, not more accurate point predictions. The last two controls reuse known conditions post hoc; all failures are retained.

Certificate / enclosure	Completed	Useful
Packed+RT (frozen comparison)	62	36
Shared-response graph (frozen comparison)	71	66
Packed, stronger centered utility bound	62	36
Same shared graph, natural-only enclosure	51	36

The 30 additional useful conditions span all 12 teachers and both designs: 18 occur at $b = 20$ and 12 at $b = 24$; both packages achieve $24/24$ at $b = 40$. Across all saved certificate thresholds in this fixed core, the shared profile is never below the controls (Appendix U); this is neither per-instance nor general dominance. The matched natural-only ablation fixes the candidate, reference, box, pivots and risk formula: disabling only centered enclosure intersections reduces useful certificates from 66 to 36. It isolates enclosure tightening, not signed cancellation alone. Rerunning packed search at 100 rather than 50 digits leaves all completion and usefulness decisions unchanged on the same 82 conditions, including stress and controls (Appendix U.3); the core remains $62/72$ complete and $36/72$ useful.

The initial six 24-leaf shared-response stress runs all trigger the original cooperative check; a subsequently frozen natural-first determinant screen completes all six usefully under matched cooperative checks without changing numerical acceptance criteria. This is a post hoc implementation improvement on three teachers, not demonstrated general scalability. All three flat-model controls refuse unique-strict certification but have useful packed bounds; both routes refuse repeated inputs. Separate 180-digit calculations check accepted parameters, determinants and prediction bounds, not the full-box theorem itself. Appendices R and U give full outcomes, certificate-threshold profiles and the configurations actually evaluated.

Targeted implementation checks. Two new post-hoc development checks address specific implementation risks (Appendix T). Across three existing models in four dyadic input units, query-only normalization yields 12/12 certificates versus 6/12 for the raw proposer, with unchanged observations and checker conditions on the same transformed prediction domain. A fixed three-response collision with one additional input exercises the complete determinant-refusal stage for ambiguous and incompatible proposals; a narrow true-proposal control certifies. These are units-robustness and checker-path tests, not new independent models, float32 guarantees, or a repair of the previous failed attack-coverage gate.

Table 2: Matched oracle access, different query budgets. Medians over 12 paired cases per size. Local CPU diagnostics, not isolated-server benchmarks. Dense is faster here.

n	NCQ responses	Dense responses	NCQ ms	Dense ms
8	4	17	0.791	0.308
32	4	65	4.730	0.877
128	4	257	49.393	4.230

Table 3: ActivitySim unknown-model certificates on $|x_i - x_i^0| \leq 0.25$ and every nonempty availability set. Unresolved cases are not accurate predictions.

Output contract	Responses	Completed	Common TV bound
Double	4	3/3	$\leq 4.28 \times 10^{-5}$
Float32	4	1/3	0.871; two unresolved
Double, preconditioned	17	3/3	$\leq 2.17 \times 10^{-6}$
Float32, preconditioned	17	3/3	0.509–0.549

5.2 A SAME-ACCESS CLASSICAL CONTROL

On 36 paired synthetic teachers ($n = 8, 32, 128$, balanced/comb, three seeds, positive/mixed-sign unknown slopes), double-precision NCQ recovers all topologies and legal parameters; maximum held-out TV is 1.13×10^{-13}. The dense control uses $2n + 1$ central-difference queries and obtains unknown scales from Jacobian reciprocity, not teacher information. Both use the same probability primitive and pass the matched 10^{-3} gate.

The query saving matters only when additional evaluations are sufficiently costly; it is not measured runtime superiority. The float32 point estimator passes 35/36 cases, rising to 36/36 with 17 responses, but this is not certification. Twelve constructed three-response collisions ($n = 3, 4, 6$) have transcript TV at most 1.943×10^{-16} but held-out differences of 0.00572–0.10131.

An identical-transcript structure fitter. We also give a conventional numerical structure fitter the *same four vectors* as NCQ. On six controlled four-alternative trees (four regular and two with one small inverse-temperature gap), it enumerates all 15 labeled binary structures and uses three predeclared least-squares starts per structure. Both methods recover all six trees, with evaluated held-out TV below 10^{-12}. These examples show no predictive-accuracy advantage for NCQ. Its contribution is a closed recovery algorithm with a sharp exact-query boundary, not exclusive access to the information in four vectors. The discrete enumeration is complete, but the continuous fits are local and supply no compatible-set certificate (Appendix O).

Sampling remains a failure. A calibrated frequency experiment on 18 teachers uses 17 inputs and a fixed half-count correction. Exact probabilities recover all 18; at $10^4, 10^6, 10^8$ draws *per input*, only 0, 2, and 10 cases recover the correct tree. Even 10^8 fails the prespecified 12/18 gate. There is no statistical sample-efficiency improvement.

5.3 PUBLIC FITTED COMPONENTS

A Swissmetro example uses public rounded Biogeme coefficients at three synthetic profiles (Bierlaire, 2026); both NCQ and dense pass, without refitting individual-level data. ActivitySim v1.3.1 supplies a 21-option component with ten internal nests and three-/five-way branching (ActivitySim contributors, 2026). Its three artificial profiles are one shared fitted component, not three datasets. The adapter exposes independently adjustable leaf-utility coordinates with synthetic diagonal input scales, preserving the upstream probability calculation. It does not recover the original travel-variable utility map or perform a whole-system transport simulation or causal policy evaluation.

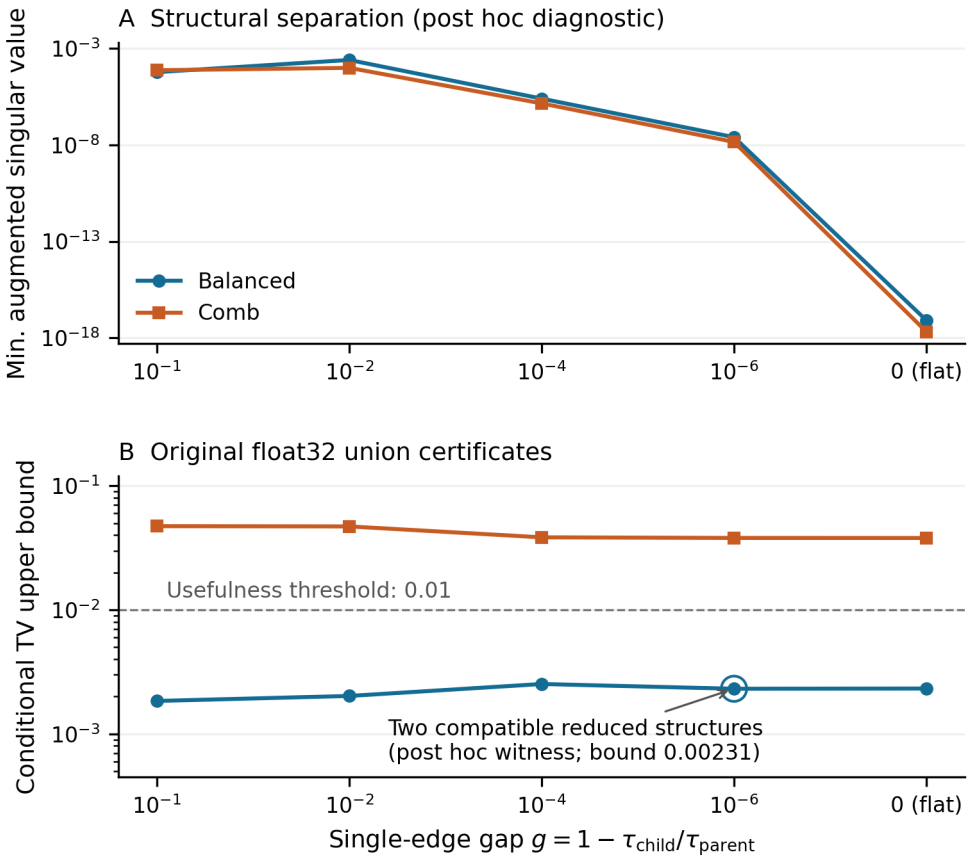


Figure 1: A single edge approaches collapse at fixed inputs and $t_{\min} = 0.4$, $n = 8$. Top: false-pair separation over all true frontiers, measured post hoc; true two-column conditioning stays near 0.0238. Bottom: original float32 common-reference bounds. The marked balanced case admits two explicitly different compatible reduced models and still meets the 0.01 gate; the comb does not. These are error-band witnesses, not bit-identical IEEE outputs.

The four-response double grammar has 97 reachable subsets and 222 splits representing 1,488,375 syntactic refinements, not validated distinct models. The uniform guarantee comes from the theorem, not the 129 input–menu checks per profile. The 17-response float32 extension fails the 0.01 usefulness gate. Separate numerical point-fit repair cannot replace an unknown-compatible-model certificate.

5.4 Actual structural ambiguity

A fixed one-edge test varies $g = 1 - \tau_{\text{child}}/\tau_{\text{parent}}$ over $0.1, 0.01, 10^{-4}, 10^{-6}, 0$ in two eight-option trees, fixing $t_{\min} = 0.4$ and four inputs. All 20 format–model runs complete. Double bounds are below 1.963×10^{-7}. Float32 balanced bounds are 0.00184–0.00253; comb bounds are 0.03795–0.04727, never useful. At $g = 10^{-6}$ the comb point fit has a wrong topology and violates temperature order despite grid TV below 1.95×10^{-7}.

The frozen second-model test holding biases unchanged fails at every positive gap. A separately labeled *post hoc* baseline-preserving bias construction supplies different reduced hierarchies compatible with all four original error intervals at $g = 10^{-6}$. The unchanged balanced union certificate is 0.002310; the comb remains 0.037974. This is feasible-model ambiguity, not merely grammar multiplicity, and not bit-identical float32 outputs. Appendix M details the witnesses and all-frontier diagnostic. They are correlated development examples, not independent confirmation.

5.5 DEVELOPMENT LADDER AND COMPLEMENTARITY

Before the fresh-model validation, a 65-condition ladder reused four ordinary strict teachers, one near-collapse teacher, three ActivitySim profiles and a zero-design control at declared budgets $b = 12, 16, 20, 24, 28, 32, 40, 48$. Packed search completes 47 conditions and refuses 18; its original 31 useful bounds become 32 with the same-reference relative-temperature refinement. The shared graph completes 31 conditions, 27 useful: six are newly useful relative to packed+RT, but all 24 non-binary public-profile conditions refuse unique-strict certification. A layerwise-error precursor failed its frozen gate and is retained, not replaced by its successful successor. These reused conditions motivated the method; they are not independent evidence beyond the fresh validation. Appendices N, P, and Q retain every budget, failure and precursor, including the zero-design refusal at $b = 48$. The ladder does not repair actual float32 failures or demonstrate better point predictions.

6 LIMITATIONS AND CONCLUSION

The oracle contract requires independently adjustable continuous inputs, full positive probabilities, nonzero affine slopes, and a promised nested-logit family. Real interfaces may restrict interventions or violate this family. A full-probability callback alone does not establish the item-wise affine-input promise; arbitrary coupled feature maps are outside this theorem. Access to the full configuration and input map may instead permit direct model copying without recovery queries. We supply no membership test, misspecification guarantee, unknown-noise interval calibration, or noisy-sample bound. Exact uniqueness holds for each fixed target almost surely, not for one deterministic design identifying all targets. Packing is worst-case exponential; local conditioning and temperature lower bounds can be poor. Public fitted-component compatibility is not new behavioral prediction accuracy. The CPU audits establish no broad neural-network or GPU performance advantage.

Within this scope, scale-free coupling reuses the same four responses through an unknown hierarchy and gives a sharp full-model boundary even against adaptive queries. With inexact data, candidate-clade checks can exclude all competing strict structures; when unique structure cannot be certified, a sound cover can still support common predictions across all available-item sets. Either route may explicitly refuse. Understanding how much of this advantage survives realistic access and uncertainty constraints remains the central open direction.

AI USE STATEMENT

AI systems were used substantially for literature search, research planning, formulation and checking of propositions and proofs, experimental design, synthetic-model code, implementation, data processing, analysis, translation, and drafting. This was not limited to language polishing. Same-workflow adversarial checks are not independent expert review; proofs have not been machine-verified. Scientific responsibility for claims, citations, experiments, and submitted text remains with the author(s).

ETHICS STATEMENT

Experiments use synthetic queries and public fitted components, not newly collected individual-level choices. No behavioral or causal deployment benefit is asserted. Prediction-interface identification may enable extraction; applications should respect permissions and access terms. Failures and limitations are retained alongside favorable results.

REPRODUCIBILITY STATEMENT

Appendices provide proofs, model definitions, fixed query seeds, interval contracts, resource caps, and disclosed corrections. Counts refer to versioned reports; overlapping audits are not independent datasets. Local timings are diagnostic. The companion package's START_HERE guide maps claims and tables to proofs and saved evidence, and separates stored-result checks from optional core inference replay; these are different reproducibility activities.

REFERENCES

Youssef M. Aboutaleb, Moshe Ben-Akiva, and Patrick Jaillet. Learning structure in nested logit models, 2020. URL https://arxiv.org/abs/2008.08048. arXiv:2008.08048.

ActivitySim contributors. ActivitySim v1.3.1: trip mode choice component and prototype-MTC configuration, 2026. URL https://activitysim.github.io/activitysim/v1.3.1/dev-guide/components/trip_mode_choice.html. Software and documentation. Accessed September 16, 2026.

Austin R. Benson, Ravi Kumar, and Andrew Tomkins. On the relevance of irrelevant alternatives. In *Proceedings of WWW*, 2016. doi: 10.1145/2872427.2883025.

Michel Bierlaire. Nested logit model, Swissmetro example b09. Biogeme documentation, 2026. URL https://biogeme.epfl.ch/sphinx/auto_examples/swissmetro/plot_b09_nested.html. Accessed September 16, 2026.

Flavio Chierichetti, Mirko Giacchini, Ravi Kumar, Silvio Lattanzi, Alessandro Panconesi, Erasmo Tani, and Andrew Tomkins. Learning multinomial logits in $o(n \log n)$ time, 2026. URL https://arxiv.org/abs/2601.04423. arXiv:2601.04423.

Matthew Kovach and Gerelt Tserenjigmid. Behavioral foundations of nested stochastic choice and nested logit, 2022. URL https://arxiv.org/abs/2112.07155. arXiv:2112.07155, version 2.

Jimmy Q. Li, Paat Rusmevichientong, Duncan Simester, John N. Tsitsiklis, and Spyros I. Zoumpoulis. The value of field experiments. *Management Science*, 61(7):1722–1740, 2015. doi: 10.1287/mnsc.2014.2066.

Matthieu Martin, Panayotis Mertikopoulos, Thibaud Rahier, and Houssam Zenati. Nested bandits. In *Proceedings of ICML*, volume 162 of *PMLR*, pp. 15093–15121, 2022. URL https://proceedings.mlr.press/v162/martin22a.html.

Panayotis Mertikopoulos and William H. Sandholm. Nested replicator dynamics, nested logit choice, and similarity-based learning, 2024. URL https://arxiv.org/abs/2407.17815. arXiv:2407.17815.

Boris S. Mityagin. The zero set of a real analytic function, 2015. URL https://arxiv.org/abs/1512.07276. arXiv:1512.07276.

Pravin Nair. Softmax is $1/2$-lipschitz: A tight bound across all ℓ_p norms, 2025. URL https://arxiv.org/abs/2510.23012. arXiv:2510.23012.

Python Software Foundation. Decimal: fixed-point and floating-point arithmetic, 2026. URL https://docs.python.org/3/library/decimal.html. Python documentation. Accessed September 16, 2026.

Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010. URL https://www.tuhh.de/ti3/paper/rump/Ru10.pdf.

Kenneth Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, second edition, 2009. URL https://eml.berkeley.edu/books/choice2.html. Chapter 4.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *25th USENIX Security Symposium*, pp. 601–618, 2016. URL https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer.

Xintong Yu, Will Ma, and Michael Zhao. Experimental assortments for choice estimation and nest identification, 2026. URL https://arxiv.org/abs/2602.16137. arXiv:2602.16137, version 2.

Yang Yu. Relative N-level operators and micro-variation coupling theory. Zenodo, 2025. URL https://doi.org/10.5281/zenodo.15736769. Version 1.3. Preprint.

A INFORMATION MODEL AND RELATION TO EARLIER WORK

We first use an unordered rooted binary tree T with labeled leaves $[n]$, $n \geq 3$, to expose the contraction; Theorem 8 extends it to arbitrary finite arity. The inclusive utilities are

$$f_i(x) = b_i + a_i x_i, \qquad f_v(x) = \frac{1}{k_v} \log\Big(e^{k_v f_A(x)} + e^{k_v f_B(x)}\Big), \tag{5}$$

where A, B are the children of v, all $k_v > 0$, and all $a_i \neq 0$. The tree is *reduced* if $k_w \neq k_v$ whenever internal w is a child of internal v. The strictly random-utility-compatible subclass used for the lower bound satisfies $k_w > k_v$ on every such edge. This is the inverse-scale form of the standard nested-logit restriction (Train, 2009). The upper bound needs only reduction, not this order.

For every nonroot vertex A define

$$s_A(x) = \frac{e^{k_{\mathrm{pa}(A)} f_A(x)}}{\sum_{B:\mathrm{pa}(B)=\mathrm{pa}(A)} e^{k_{\mathrm{pa}(A)} f_B(x)}}.$$

The probability of leaf i is the product of s_A along its root-to-leaf path. For any vertex A, write $P_A(x) = \sum_{i \in A} p_i(x)$, where membership means descent from A. Then $P_{\mathrm{root}} = 1$ and $P_A / P_{\mathrm{pa}(A)} = s_A$. Equivalently, p_i is the derivative of the root utility with respect to the additive *latent leaf utility*; the physical derivative is $\partial_{x_i} f_{\mathrm{root}} = a_i p_i$, which is different when $a_i \neq 1$.

Oracle and target. A query specifies $x \in \mathbb{R}^n$ and observes the full positive vector $p(x)$. Queries may lie in an arbitrarily small open box; there is no restriction on depth. A nonadaptive design selects all query inputs independently of the unknown model and before seeing responses. Recovery means identifying the labeled reduced tree and the function $p(\cdot)$ on all of $\mathbb{R}^n$. The theorem is conditional on membership in (5); finite agreement cannot certify membership of an arbitrary black box.

Gauge. The transformation $f \mapsto sf + c$, $a \mapsto sa$, $b \mapsto sb + c$, $k \mapsto k/s$, $s > 0$, preserves $p(\cdot)$. We use $k_{\mathrm{root}} = 1$ and $b_n = 0$ as the canonical gauge when counting parameters or assigning a prior. Each tree then has $d = 3n - 3$ continuous parameters: n slopes, $n - 1$ biases, and $n - 2$ nonroot temperatures. For reconstruction alone, $f_{\mathrm{root}}(x^0) = 0$ is a more convenient shift convention, converted to $b_n = 0$ at the end.

Closest identification results. Benson, Kumar, and Tomkins (Benson et al., 2016) reconstruct nests from an oracle expressing IIA relations between triples and connect it to choice data. Kovach and Tserenjigmid (Kovach & Tserenjigmid, 2022) derive behavioral characterizations and structure identification through probability-ratio restrictions. Aboutaleb, Ben-Akiva, and Jaillet (Aboutaleb et al., 2020) jointly optimize nesting structure and parameters from choice observations using mixed-integer nonlinear optimization. These works establish that learning nests, using ratios, and reasoning about trees are not new by themselves.

The closest recent work is Yu, Ma, and Zhao (Yu et al., 2026). Their main nest-partition result uses $O(\log n)$ availability experiments, with a matching lower bound, and their multilevel extension already uses sibling ratios and recursive identification. Our model permits continuous item-wise interventions with unknown affine scales while keeping every item available. The contribution is a complete all-depth contraction and exact full-model query boundary under that different permission. Four numerical-input responses do not contradict a lower bound for binary availability experiments. Moreover, our negative frequency experiments do not match their statistical and deployment evidence.

Continuous interventions also have a long history in demand learning. Li et al. (Li et al., 2015) use randomized price experiments to estimate cross-elasticity structure under sparsity and influence assumptions. We do not assume sparse interactions, but impose the restrictive nonlinear hierarchy (5) and exact probability access. Neither information contract subsumes the other's practical guarantee. The present experiment is not a comparison against state-of-the-art observational choice estimation.

Prediction-score access is also central to model extraction. Tramèr et al. (Tramèr et al., 2016) show how real-valued confidence outputs support equation-solving recovery of several model classes. Thus recovering parameters from probability equations is not itself novel. Our specific problem

combines an unknown deep nesting topology with unknown item-wise input units, and asks for the exact query optimum for its full probability function, even allowing adaptive designs. Our experiments use only controlled local models and do not claim an attack on a deployed service.

Our stability argument uses the established nested-entropy representation developed in nested bandits and similarity-based learning (Martin et al., 2022; Mertikopoulos & Sandholm, 2024). Their forward power-sum updates preclude claiming finite message transport alone as new. The unknown topology, unknown input units, and query boundary distinguish the present task. Softmax norm bounds (Nair, 2025) are related ingredients, not results that this paper replaces. Recent MNL learning bounds from sampled winners (Chierichetti et al., 2026) also concern a different observation oracle; four probability vectors do not contradict those sample-complexity bounds.

B SCALE-FREE FINITE COUPLING

Fix a baseline x^0 and let $\Delta g(x) = g(x) - g(x^0)$. An active frontier consists of disjoint true subtrees that cover all leaves. Initially each leaf is active; contracting true siblings preserves this property. Maintain a feature

$$u_A(x) = \begin{cases} x_i - x_i^0, & A = i \text{ is an original leaf,} \\ k_A \Delta f_A(x), & A \text{ is an internal active vertex.} \end{cases} \tag{6}$$

The internal feature is dimensionless. Its definition does not require that k_A be separately known.

Lemma 4 (Finite coupling and contraction)**.** If active A, B are siblings with parent C, then

$$z_{AB}(x) := \Delta \log \frac{P_A(x)}{P_B(x)} = c_A u_A(x) - c_B u_B(x), \tag{7}$$

$$c_A = \begin{cases} k_C a_i, & A = i, \\ k_C / k_A, & A \text{ internal}, \end{cases} \qquad u_C(x) = \log\Big(w_A e^{c_A u_A(x)} + w_B e^{c_B u_B(x)} \Big), \tag{8}$$

$$w_A = \frac{P_A(x^0)}{P_A(x^0) + P_B(x^0)}, \qquad w_B = 1 - w_A.$$

Here $u_C = k_C \Delta f_C$ and $P_C = P_A + P_B$.

Proof. Conditioning on the common parent cancels all ancestors: $\log(P_A/P_B) = k_C(f_A - f_B)$. Taking a baseline difference gives (7). Dividing the two log-sum-exp expressions defining $e^{k_C f_C(x)}$ and $e^{k_C f_C(x^0)}$ gives (8). □

Equation (7) has two unknown coefficients, not an unknown function. An internal-child coefficient is positive; an original leaf coefficient can be signed. The contraction uses only these coefficients and observed baseline masses. It therefore removes a potential circular dependency: one does not need the parent's absolute temperature before constructing its message.

When utility units are already known, the same identity is a finite power-law coupling. With $X = e^{f_A - f_B}$ and $Y = P_A/P_B$, it reads $Y = X^{k_C}$; a constant multiplier can be absorbed by the baseline difference. This connects the construction to the finite-comparison and power-law case of the original micro-variation framework (Yu, 2025). The additional operation here is transporting the identifiable relation through a hierarchy with unknown scales.

B.1 WHY FALSE SIBLINGS DO NOT SATISFY THE SAME RELATION

We use a derivative identity only to prove identifiability; the algorithm does not request derivatives. For a frontier, regard its inclusive utilities as independent variables. Their normalized cross derivatives have labels λ_v such that

$$\frac{\partial_{f_A} \partial_{f_B} f_{\mathrm{root}}}{P_A P_B} = \lambda_{\mathrm{lca}(A,B)}, \qquad \lambda_{\mathrm{root}} = -k_{\mathrm{root}}, \qquad \lambda_w - \lambda_v = \frac{k_v - k_w}{P_w} \tag{9}$$

for an internal child w of v. For completeness, write the composition outside v as $G(f_v)$ with $G' = P_v$. Mixed differentiation between different children gives $\lambda_v = G''/P_v^2 - k_v/P_v$. Putting

$s = P_w/P_v$, differentiation within a child gives

$$\partial^2_{f_w} f_{\text{root}} = G'' s^2 + P_v k_v s(1-s).$$

Substitution in the corresponding expression for λ_w proves the recursion in (9); descendants inherit the same label through the chain rule.

Lemma 5 (False-sibling separation)**.** At every true frontier, if A, B are not siblings, the three real-analytic functions u_A, u_B, z_{AB} are linearly independent on every open input box.

Proof. Let $v = \text{lca}(A, B)$. At least one child w of v, say the one containing A, lies strictly above A. Choose an active vertex C in the other child branch of w from A. It is disjoint from both A and B. Shifting every latent leaf utility in C by δ shifts f_C by δ, leaving f_A, f_B unchanged; the nonzero a_i make this a valid input perturbation. By (9),

$$\partial_{f_C} \log(P_A/P_B) = P_C(\lambda_w - \lambda_v) = P_C \frac{k_v - k_w}{P_w} \neq 0. \tag{10}$$

The baseline term is constant. Thus z_{AB} cannot be a fixed linear combination of u_A, u_B. Those two features are themselves linearly independent because they depend on disjoint coordinates and can be varied independently. The nonzero label difference is between *adjacent* internal vertices; nonadjacent labels need not be distinct. □

Lemma 6 (Generic evaluation rank)**.** If $g_1, \ldots, g_m$ are linearly independent real-analytic functions on a connected open box, their evaluations at m independent absolutely continuous random inputs form an invertible matrix almost surely.

Proof. The evaluation vectors span $\mathbb{R}^m$; otherwise a nonzero vector orthogonal to their span gives a constant linear dependence of the functions. Hence their evaluation determinant is a nonzero analytic function of the m inputs. Its zero set has Lebesgue measure zero (Mityagin, 2015). □

C Recovery from four responses

Theorem 7 (Four-response identification)**.** Fix any reduced model (5) and any baseline x^0. The response at x^0 and responses at three independent absolutely continuous inputs in an open box identify its tree and its entire probability function almost surely. The query design is nonadaptive and does not depend on n, depth, hidden temperatures, or input scales, except for its input dimension.

Algorithm: finite-coupling contraction. Maintain the four observed masses and the three non-baseline features for each active vertex. For a candidate pair, form the 3×2 matrix with columns $u_A, -u_B$ and vector z_{AB}. Accept the pair if this matrix has rank two, its linear system is exactly consistent, its internal-child coefficients are positive, and its leaf coefficients are nonzero. Replace an accepted pair by (8), save its coefficients, and repeat until the root is reached. If no pair passes or final parameters are illegal, report failure. The proof shows that neither failure occurs for the promised model outside a null set of query designs.

Proof. For a true pair, the columns u_A, u_B are independent functions; Lemma 6 makes their three evaluations rank two almost surely. The unique fitted coefficients are the legal coefficients of Lemma 4. For a false pair, Lemmas 5 and 6 make the augmented 3×3 matrix nonsingular almost surely, so its two-coefficient system is inconsistent.

Fix a true contraction path. There are finitely many frontier tests along it, hence the union of all exceptional designs has measure zero. There are also finitely many true contraction paths. Taking the union over all of them permits any ordering among passing pairs, without falsely assuming that observations remain independent after conditioning on earlier selected merges. Outside this union only true siblings pass, and Lemma 4 preserves the feature invariant. Every nontrivial binary tree has a cherry, so induction reaches the root.

Set $k_{\text{root}} = 1$. For every saved edge $C \to A$, recover

$$k_A = k_C/c_A \quad (A \text{ internal}), \qquad a_i = c_i/k_C \quad (A = i). \tag{11}$$

Set $f_{\mathrm{root}}(x^0) = 0$ and descend using

$$f_A(x^0) = f_C(x^0) + \frac{\log(P_A(x^0)/P_C(x^0))}{k_C}. \tag{12}$$

Finally let $b_i = f_i(x^0) - a_i x_i^0$. Equations (11)–(12) reconstruct the original parameters up to the common positive scale and common additive shift. They therefore recover $p(x)$ for every x, not only the queried values. Subtracting the resulting b_n from all biases gives the canonical gauge. □

Exact computational count. Test each initial pair once. After a merge, only pairs involving the newly created vertex need testing, since all other messages and masses are unchanged. The total number of constant-size tests is $\binom{n}{2} + \sum_{j=0}^{n-2} j = (n-1)^2$. An ideal exact implementation stores only passing pairs. These are true cherries and there are $n-1$ such merges over the run. Thus it needs $O(n^2)$ arithmetic operations and $O(n)$ stored real numbers and tree pointers. This counts log/exp and exact equality as unit operations; it is not a bit-complexity claim. Our numerical implementation instead retains a minimum-residual heap containing failed as well as passing candidates and has $O(n^2)$ storage.

Theorem 8 (Arbitrary finite arity)**.** The four-response conclusion holds for every reduced finite tree with internal arity at least two. Binary contraction followed by flattening equal-temperature edges recovers its reduced hierarchy and full probability function.

Proof. An active group may now be a union of whole children of one original vertex. At that vertex's temperature its inclusive value is a log-sum-exp over those children. Grouping two exposed children preserves Lemma 4, including coefficient one on an artificial equal-temperature edge. Collapse active groups to leaves and flatten equal-temperature edges in the remaining effective hierarchy. An incompatible pair crosses a reduced internal edge; the softmax derivative calculation in (9) holds at any arity, so (10) again supplies an external varying group and a nonzero analytic determinant. A finite union over compatible grouping paths excludes all false merges almost surely. There is always an exposed compatible pair, so contraction finishes. Reconstruct parameters and flatten equal edges: precisely the introduced refinements disappear, while original reduced edges remain. □

The arithmetic-test count remains $(n-1)^2$, but the binary linear-memory argument does not extend: a star has $\binom{n}{2}$ passing pairs. The finite-arity cached version uses $O(n^2)$ memory. These ideal counts are separate from the interval search below.

Corollary 9 (Calibrated sufficient count)**.** If all $a_i = 1$ are known, a baseline and two independent continuous perturbations suffice almost surely.

Proof. Use $u_A = \Delta f_A$ at every vertex, starting with the known leaf changes. True siblings satisfy $z_{AB} = k_C(u_A - u_B)$, with only one unknown coefficient. For a false pair, varying the external C in (10) changes z_{AB} but not $u_A - u_B$. Two generic evaluations distinguish the two functions. After estimating k_C, divide (8) by k_C to form Δf_C. The finite-path argument is unchanged. □

We do not claim that three is the optimal calibrated count.

D WHY THREE NONADAPTIVE RESPONSES CANNOT SUFFICE

The unknown-scale family has $3n - 3$ continuous parameters, exactly the number of independent coordinates in three response vectors. This observation alone does *not* prove a lower bound. In fact, each fixed tree can be locally identified from three suitably chosen responses. The obstruction is a combination of unresolved continuous fibers for degenerate designs and overlapping response images of different trees for nondegenerate designs.

Lemma 10 (Prescribed-tree local inverse)**.** Fix T and inputs x^0, x^1, x^2. Given positive response vectors, solve (7) as a 2×2 system at each prescribed merge and apply (8). Where all these systems are invertible and all recovered k_v, a_i are positive, this construction and (11)–(12) give an analytic local inverse of the forward response map. The statement temporarily permits either temperature order, including equality on an edge.

Proof. The square systems uniquely determine their coefficients. All subsequent operations are analytic while determinants, masses, coefficients and temperatures stay in their stated open domains. The recovered child ratio equals the prescribed observed ratio at each of the three inputs. Together with $P_C = P_A + P_B$, this fixes both conditional probabilities at that merge. Inducting down the tree gives the prescribed leaf probabilities. Conversely, data from parameters with nonsingular designs return their original coefficients, temperatures, slopes and biases in the fixed gauge. Thus the maps are analytic two-sided inverses on neighborhoods. The response coordinates can be taken as $y_{t,i} = \log(p_i(x^t)/p_n(x^t))$, $t = 0, 1, 2$, $i < n$, so both spaces have dimension $d = 3n - 3$. □

Theorem 11 (Nonadaptive lower bound)**.** For $n \geq 3$, no measurable randomized nonadaptive algorithm with at most three full probability responses recovers $p(\cdot)$ with probability one for every model (5) with $a_i > 0$ and strict order $k_w > k_{\mathrm{pa}(w)}$. This holds even though all probabilities are exact.

Proof outline. For a fixed three-input design, a dependent pair of leaf perturbation vectors leaves a continuous parameter fiber in a topology containing that cherry. Otherwise three binary refinements of a common ternary boundary admit regular local inverse charts, and two strict-RUM response images overlap on an open set. A proper full-support prior fixed before all query coins gives positive Bayes error for every fixed design. Integrating over randomness yields a fixed hard target. Appendix F supplies the full construction and quantifier argument. □

Corollary 12 (Exact nonadaptive optimum)**.** Four is the optimal number of full responses for probability-one, nonadaptive exact recovery of the entire unknown-scale model within the strict compatible family.

The lower bound asserts no uniform numerical failure probability. It does not establish a three-response lower bound for topology alone in every design. Case 1 is a parameter ambiguity; Case 2 also provides different-topology witnesses.

Theorem 13 (Adaptive boundary for every size)**.** For each fixed $n \geq 3$, no measurable randomized adaptive policy using at most three finite input vectors recovers the entire model almost surely for every strict-RUM binary model with positive unknown input scales.

Appendix G conditions successively on earlier responses and constructs three regular inverse charts in the same conditional parameter fiber. It does not apply a fixed-design proof after adaptation or pad a three-item instance. Combined with Theorem 7, four responses are optimal for each $n \geq 3$ even allowing adaptive queries.

E FINITE PRECISION: OUTER COVERS AND HONEST REFUSAL

Exact continuous observations can distinguish values that become indistinguishable after quantization or sampling. An elementary representation bound illustrates why relative precision and absolute truncation should be distinguished.

Proposition 14 (Log-ratio representation error)**.** Suppose returned positive numbers satisfy $|\widehat{p}_i/p_i - 1| \leq \eta < 1$, before any optional common normalization. For any two disjoint nonempty subsets A, B, the error in $\log(P_A/P_B)$ is at most $2\,\mathrm{atanh}\,\eta$. The error in its baseline difference is at most $4\,\mathrm{atanh}\,\eta$.

Proof. Every subset sum lies between $(1 - \eta)P_A$ and $(1 + \eta)P_A$. The ratio's multiplicative error therefore lies between $(1 - \eta)/(1 + \eta)$ and its reciprocal. Taking logarithms and then a difference proves the bounds. Common normalization cancels in the ratio. □

This standard calculation is not itself an end-to-end recovery certificate. Even if raw log-ratio errors are small, a sibling design can have an arbitrarily small least singular value, a false sibling can have an arbitrarily small residual, and reconstructed messages propagate earlier error. The exact theorem supplies no uniform lower bounds on these quantities. Likewise, a very small leaf probability can round to zero under a fixed-decimal interface. Normal float32 rounding of exact positive probabilities can satisfy a relative-error bound; underflow, approximate numerical model evaluation, and

empirical choice frequencies need not satisfy that premise. Adding a hidden pseudocount changes the information model rather than repairing the theorem.

Finite probabilities also cannot certify that a black box globally belongs to the family. A smooth perturbation vanishing at the queried inputs can preserve every observed response while changing unqueried responses. Membership, cross-nesting, mixtures, nonlinear input calibrations, and shared input coordinates therefore require separate assumptions or tests. Our global reconstruction guarantee is a promised-family result.

E.1 From output intervals to compatible models

Suppose a finite transcript supplies positive intervals I_{ti} containing every true $p_i(x^t)$. For this certificate we restrict to weak RUM order: with $\tau_v = 1/k_v$, internal children satisfy $0 < \tau_A \le \tau_C$; unknown leaf slopes may have either sign. We fix $\tau_{\mathrm{root}} = 1$ and $f_{\mathrm{root}}(x^0) = 0$. All binary refinements of a multiway RUM tree are admitted.

For each candidate pair, evaluate mass sums, baseline log ratios, and features using outward intervals. Invert each two-row design whose interval determinant excludes zero, and intersect its coefficient boxes. Prune only on an empty intersection, an internal coefficient disjoint from $(0, 1]$, or a residual interval excluding zero. Apply (8) by directed arithmetic using both observed baseline-weight intervals. A singular interval inverse is an unresolved possibility, not an incompatible pair.

A greedy choice is unsafe. Preserve all partial binary trees through one interval hull U_M for each retained leaf subset M, and all accepted splits $M = A \dot\cup B$. Processing subsets by increasing cardinality makes child hulls complete before they are used at a parent. This is a directed acyclic grammar of candidate trees. A resource limit or any opaque branch causes global refusal in our conservative prototype. On its root-reachable part, descend in decreasing cardinality and unite parameter images from all possible parents:

$$\tau_A = c_A \tau_C, \quad a_i = c_i \tau_C, \quad f_A^0 = f_C^0 + \tau_C \log(P_A^0 / P_C^0), \quad b_i = f_i^0 - a_i x_i^0.$$

Theorem 15 (Conditional compatible-model outer cover)**.** Assume the intervals contain a transcript from the promised weak-RUM class. If the search closes without an opaque inverse or a resource refusal, every compatible binary tree is generated by its grammar, and its normalized parameters lie in the corresponding propagated hulls.

Proof. Induct on a compatible model's subtree size. Leaves have exact input-change features. When child messages are enclosed, every safely inverted local equation contains the true coefficients. Intersection, ratio-order checks, and residual exclusion cannot remove that solution. Directed contraction encloses the parent message. A hull over splits only enlarges the cover, so all true splits survive to the full set. In the downward pass, all parents precede their children by cardinality. Unioning their images encloses true temperatures and baseline potentials, hence slopes and biases. Losing correlations can introduce spurious possibilities but cannot remove the true model. □

This is a necessary outer cover, not proof that every generated tree or box point is feasible. Membership and observation-error authority are assumptions, not outputs of four-query agreement. At most $2^n - 1$ subset states may occur, so worst-case search remains exponential. Exact polytomies cannot generally be distinguished from arbitrarily small true gaps at nonzero precision on a bounded input domain: continuity gives overlapping relative-response bands. Ambiguity should be retained, not hidden by a collapse tolerance.

E.2 Predictions across uncertain structures

A common prediction bound requires more than a before/after repair comparison. First compare two positive-temperature models on one finite tree, with leaf-utility vectors u, u'. For any nonempty available set S, remove missing alternatives and empty branches; unary vertices copy their child. Put

$$E = \tfrac{1}{2}\{\max_i(u_i' - u_i) - \min_i(u_i' - u_i)\}, \quad \delta = \max_v |\tau_v' - \tau_v|, \quad t_{\min} = \min_v \min(\tau_v, \tau_v') > 0.$$

Theorem 16 (All-assortment temperature stability)**.** For every such S, with $C_n = \sqrt{(\log n)^2 + 2\log n + 2}$,

$$\mathrm{TV}(p'_S, p_S) \le \min\left\{1, \frac{E + \delta C_n}{2t_{\min}}\right\}. \tag{13}$$

The bound has no explicit depth factor; conditioning through $t_{\min}$ remains. The order $\log n$ cannot generally be removed from a uniform linear temperature bound.

Appendix H uses the classical nested-entropy representation: curvature telescopes to a leaf Fisher metric, and temperature perturbations are bounded by leaf surprisal. This avoids multiplying layer-wise Lipschitz constants, but is not a depth-independent guarantee for a family with vanishing $t_{\min}$.

Intersect the clades of all grammar trees to obtain their common hierarchy. Choose positive, weakly ordered temperatures and one slope/bias vector on it. In each binary derivation, assign every inserted node its nearest common ancestor's temperature. Equal-temperature flattening yields the same reference function on every derivation. Let e bound all leaf-utility deviations from this reference on a declared input box, let δ bound all node-temperature deviations, and let $t_{\min}$ bound all temperatures below. Applying (13) with $E \le e$ certifies every compatible structure, every input in the box, and every nonempty S simultaneously. Appendix I gives the combinatorial and union details.

No unique topology is required. Conversely, small slope error has no useful guarantee on all $\mathbb{R}^n$: slopes $(1, 1)$ versus $(1+\epsilon, 1)$ and input $(t, (1+\epsilon/2)t)$ give opposite log odds $\mp\epsilon t/2$, and TV tends to one. Uncertain slopes therefore require a bounded domain or equivalent utility-discrepancy bound. For rewards in $[0, R]$, the distribution certificate implies at most R times its bound in any fixed expected reward, or twice this for optimization over a common assortment family. This does not give an assortment solver or real-user regret.

F Full nonadaptive lower-bound proof

Proof. We first fix three input vectors, allowing repeated or degenerate ones. Write $d_i = (x_i^1 - x_i^0, x_i^2 - x_i^0) \in \mathbb{R}^2$.

Case 1: some d_i, d_j are linearly dependent. Choose a topology containing the leaf cherry $\{i, j\}$. There is a nonzero $h \in \mathbb{R}^2$ orthogonal to both vectors. For every parameter choice in this topology,

$$\sum_{t=1}^{2} h_t \left[\log \frac{p_i(x^t)}{p_j(x^t)} - \log \frac{p_i(x^0)}{p_j(x^0)}\right] = 0. \tag{14}$$

This is a nontrivial linear constraint on the d response log-odds coordinates. The analytic forward map from the d-dimensional canonical parameter domain therefore has rank at most $d - 1$ everywhere. At a point of maximal rank, a sufficiently small neighborhood has constant rank. The constant-rank theorem supplies local source coordinates (s, t) in which the response depends only on s and at least one continuous coordinate remains in t. Under any everywhere-positive continuous parameter density, a single decoded value of t has conditional probability zero on this patch. A change of variables and Fubini's theorem imply positive average failure, even if the tree is disclosed to the decoder. Different canonical parameters on the patch represent different entire functions: Theorem 7, applied to any two candidates and the union of their null sets, gives uniqueness of the canonical representation.

Case 2: every pair d_i, d_j is linearly independent. Construct a common tree with one ternary vertex with children A, B, C and strict temperature order on every other internal edge. For $n = 3$ these children can be leaves; for larger n embed this local construction in a shared tree. Consider the three binary refinements merging AB, AC or BC first. Set the introduced internal temperature equal to its parent's. By associativity of log-sum-exp at equal temperatures, the three refinements represent the same central probability function. This center is used only inside the open positive-parameter domain of Lemma 10, not as a final reduced hard instance.

There are finite leaf biases for which all local inverse designs are nonsingular in all three refinements. To prove existence, fix any one design with disjoint child leaf sets S, R, set all leaf slopes to one, and select $i \in S, j \in R$. Concentrate each child's utility on its selected leaf by sending its other leaf biases to minus infinity. Its two feature changes tend to d_i for an original leaf or $k_S d_i$ for an internal

child; the other child analogously tends to a positive multiple of d_j. The limiting determinant is nonzero. Thus this determinant is not the zero analytic function of the finite bias vector. There are finitely many determinants across the three trees, so their zero sets cannot cover the bias space. Choose one common finite bias vector avoiding all of them. A common utility shift enforces $b_n = 0$ without changing any determinant or response.

Lemma 10 now gives three analytic response charts in a common neighborhood of the central response y^*. For refinement r, let

$$h_r(y) = k_{\mathrm{new},r}(y) - k_{\mathrm{parent},r}(y).$$

Each $h_r(y^*) = 0$ and each differential $dh_r(y^*)$ is nonzero, because this temperature difference is a freely variable coordinate in that tree's parameter chart. All positive slopes and all other strict internal inequalities persist on a small common neighborhood. Among three nonzero linear functionals, some two have a common strictly positive direction. Indeed two open positive halfspaces are disjoint only if their normals are oppositely proportional; three nonzero normals cannot be pairwise oppositely proportional. Choose such a direction v. For sufficiently small $\epsilon > 0$, both selected $h_r(y^* + \epsilon v)$ are positive. An open neighborhood of this response is therefore represented by two different strictly compatible, reduced trees. Their entire probability functions differ by the uniqueness consequence of Theorem 7.

Randomized designs and decoders. Before drawing the query random coins, fix a proper prior with positive weight on every labeled binary tree and an everywhere-positive continuous density on each strict canonical parameter domain. This prior is independent of x^0 and of the queries. For every fixed design, Case 1 gives a positive-prior-mass patch with unresolved fibers, or Case 2 gives an open overlap of response images whose two induced densities are positive. In Case 2, even the optimal decoder has positive Bayes error on the overlap, since a decoded function cannot equal the two distinct functions simultaneously. Randomizing the decoder does not remove either obstruction. Thus the prior-averaged failure probability conditional on every fixed design is strictly positive. Its integral over the query coins remains positive. Fubini's theorem then gives at least one fixed target with positive failure probability for that randomized algorithm. The hard target is not chosen after observing the query seed. Fewer queries can be padded to three. □

An explicit regular collision. A concrete example illustrates the overlap without numerical fitting. Let $L = \log 2$, use leaves $1, 2, 3$, and fix the root inverse temperature to one. For tree $A = ((1, 2), 3)$ and tree $B = ((1, 3), 2)$ set

$$k_A = \frac{\log(5/32)}{\log(5/16)}, \qquad a_i^A = 2/k_A, \qquad b^A = (L(1 - 1/k_A), L(1 - 1/k_A), 0),$$

$$k_B = 1 + \frac{L}{2\log(9/8)}, \quad a_i^B = 3/(2k_B), \quad b^B = (0, -L(1 - 1/k_B), 0).$$

Both $k_A, k_B > 1$ and all slopes are positive. At inputs $x^0 = (0, 0, 0)$, $x^1 = (1, 1, 1)$ and $x^2 = (0, L, 2L)$, both return $p^0 = p^1 = (1, 1, 1)/3$ and $p^2 = (1, 4, 8)/13$. The baseline biases give uniform probabilities; the first change is a common utility shift. At x^2, tree A has within-cherry odds $1 : 4$ and group/singleton odds $5 : 8$, whereas tree B has within-cherry odds $1 : 8$ and group/singleton odds $9 : 4$. These follow from $k_A a_i^A = 2$, $k_B a_i^B = 3/2$ and the displayed logarithms.

The cherry/root inverse-design determinants are $-L, \log(5/32)$ for A and $-2L, \log(9/(4\sqrt{2}))$ for B, all nonzero. Lemma 10 and the strict inequalities thus give an open response neighborhood shared by the two families, not just an isolated collision. Their full functions differ: at $(0, 0, t)$, tree A has $p_1 = p_2$, while in tree B

$$\frac{d}{dt}\log(p_1/p_2) = a_3^B(1 - k_B)\frac{p_3}{p_1 + p_3} < 0.$$

This illustrates one fixed design; it does not replace the all-design and fixed-prior argument above.

G Proof of the adaptive boundary for every size

We prove Theorem 13 separately for every $n \geq 3$. It is not a reduction obtained by padding a three-item instance: additional alternatives can reveal additional inclusive-value information. All final

hard instances are strict-RUM binary trees with positive slopes. Collapsed edges are used only as common boundary centers. Write $\tau_v = 1/k_v$, with $\tau_{\text{root}} = 1$, and normalize $b_n = 0$. We represent recovered models by canonical parameters. A decoder is measurable into the finite disjoint union of these canonical parameter spaces, augmented by a failure symbol.

A prior fixed before all coins. Give each labeled binary topology positive mass and, within it, a normalized density proportional to

$$\exp\left(-\sum_i a_i - \sum_{v\neq\text{root}} (k_v - 1) - \sum_{i<n} b_i^2\right)$$

on its strict ordered, positive-slope domain. This density is proper and has a positive smooth extension at equal temperatures. Fix all policy and decoder coins. We show positive prior-averaged failure for the resulting deterministic measurable policy, then integrate over the coins. Distinct canonical parameters in this class give distinct full functions by Theorem 7: a design outside two fixed models' exceptional null sets would otherwise uniquely reconstruct both.

Conditional ambiguity principle. An analytic response map on an open d-dimensional parameter patch with maximum rank $r < d$ cannot recover a parameter drawn from a positive density. On a positive-mass maximum-rank patch, constant-rank coordinates separate response coordinates from a positive-dimensional continuous fiber. Fubini gives probability zero of correctly guessing a single fiber point there. After conditioning on an earlier submersion, the same argument applies to its conditional parameter density for almost every response. For a singular earlier map, first restrict to a maximum-rank patch and use its own response coordinates, not a fictitious full-dimensional output density. Countably many local charts suffice. We never differentiate the adaptive policy: each query becomes fixed only after conditioning on preceding responses.

Baseline coordinates and the first intervention. The first input x^0 is fixed. Let its response be p^0. For any fixed temperatures, the baseline leaf utilities with root value zero are

$$v_i^0 = \sum_{v\to c \text{ on path } i} \tau_v \log(P_c^0/P_v^0). \tag{15}$$

Subtract $a_i x_i^0$ and make a common shift to set $b_n = 0$. This is an analytic inverse between baseline probability coordinates and bias differences at fixed slopes and temperatures. Consequently, for almost every interior p^0, every topology has positive conditional weight and positive density on its $2n-2$ slope and nonroot-temperature coordinates. This only reparameterizes the fixed prior. Condition on p^0 and set $t = x^1(p^0) - x^0$.

For any positive temperatures, the Hessian H of the root potential in leaf utilities is positive semidefinite with nullspace $\operatorname{span}\{\mathbf{1}\}$. Indeed recursive differentiation gives, at a vertex v,

$$z^\top H_v z = \sum_c q_{c|v} z_c^\top H_c z_c + \tau_v^{-1} \operatorname{Var}_{q_{\cdot|v}} \left(\nabla F_c \cdot z_c\right). \tag{16}$$

The first terms force z to be constant within each child; the variance forces those constants to agree. All conditional weights are positive. Thus H has rank $n-1$, and any $n-1$ columns are independent: a null vector supported on fewer than n leaves must vanish. At fixed p^0, τ, the slope derivative of p^1 is $H\operatorname{diag}(t)$, since the leaf changes are $a_i t_i$. A common gauge shift is killed by H.

If at least two coordinates t_i, t_j vanish, restrict to a positive-prior topology with cherry (i,j). It obeys $p_i^1/p_j^1 = p_i^0/p_j^0$. The first-intervention map on $2n-2$ parameters has maximum rank $r \le n-2$. On a positive-mass maximum-rank patch, conditioning on p^1 leaves dimension $2n-2-r \ge n$. The now-fixed last query has at most $n-1$ response coordinates. The conditional ambiguity principle gives positive failure, even if the topology is disclosed. This includes $t = 0$ and all singular cases of this type.

A full-rank first intervention. Suppose at least $n-1$ coordinates of t are nonzero. Equation (16) gives rank $n-1$ for the slope-to-p^1 map in every topology, also at the flat star. Choose a nonempty open set U of p^1 values attained at that star by positive slopes. Every $p^1 \in U$ is attainable in every

strict binary topology: perturb its temperatures into strict order and compensate with slopes by the implicit function theorem. There are finitely many topologies. For almost every $p^1 \in U$, each has positive conditional density on an $(n-1)$-dimensional first-two-response fiber. Fix (p^0, p^1) and then $s = x^2(p^0, p^1) - x^0$.

Write $d_i = (t_i, s_i)$. If d_i, d_j are dependent, their first coordinates are not both zero, so $(s_i, s_j) = \alpha(t_i, t_j)$ for some finite α. In the positive-support topology with cherry (i, j),

$$\log(p_i^2/p_j^2) - \log(p_i^0/p_j^0) = \alpha\{\log(p_i^1/p_j^1) - \log(p_i^0/p_j^0)\}.$$

The final response then has rank at most $n-2$ on the $(n-1)$-dimensional conditional fiber, and the ambiguity principle applies. Henceforth all leaf pairs d_i, d_j are independent.

Coordinates that preserve both preceding responses. For a prescribed tree define $\ell_i = \log(p_i^1/p_i^0)$ and

$$B_i(\tau) = \sum_{v \to c \text{ on path } i} \tau_v \log \frac{P_c^1/P_v^1}{P_c^0/P_v^0}.$$

For a root inclusive-value change D, the necessary and sufficient leaf changes are

$$a_i t_i = D + B_i(\tau). \tag{17}$$

Together with (15), these formulas reconstruct both prescribed responses by descending conditional ratios. When every $t_i \neq 0$, $(\tau_{\text{nonroot}}, D)$ are fiber coordinates. If exactly $t_j = 0$, set $D = -B_j(\tau)$ and use $(\tau_{\text{nonroot}}, a_j)$ instead. They are analytic on the legal open portions, and extend analytically to arbitrary signed slopes while all temperatures stay positive.

We construct a common center whose root has three children: singleton leaves j, b and a fixed binary subtree C containing all other leaves. For $n = 3$, C is a leaf. The three binary refinements merge jb, jC, or bC at a newly introduced node of temperature one. They represent the same function at the center. All edges inside C will be strictly ordered below the root temperature. Lemma 10 gives a local three-response inverse for each refinement whenever every two-by-two message design is nonsingular. The essential step is simultaneous regularity *without changing* p^0, p^1 *or* s.

All first-intervention coordinates are nonzero. At the completely flat center, (17) gives $a_i = (\ell_i + D)/t_i$ for D in a nonempty open legal interval. Set $\xi_i = s_i/t_i$; all ξ_i are distinct. For any internal group S in any of the three refinements,

$$u_S^1 = D + \log(P_S^1/P_S^0), \qquad u_S^2 = \log \sum_{i \in S} \frac{p_i^0}{P_S^0} e^{(\ell_i + D)\xi_i}.$$

As $D \to +\infty$, this vector is $(D + O(1), D \max_{i \in S} \xi_i + O(1))$. Disjoint groups have different extreme ξ values. An internal/internal design determinant has a nonzero quadratic leading term; a leaf/internal determinant has a nonzero linear leading term since $t_i \neq 0$ and ξ_i is distinct from the group's values. Leaf/leaf determinants are nonzero by hypothesis. Thus each is an analytic, nonidentically-zero function of D on $\mathbb{R}$. Finitely many zero sets cannot cover the open legal interval. Choose a legal D avoiding them, then perturb the C temperatures slightly from one into strict order using (17). The prefix stays fixed, all slopes stay positive, and all three charts stay regular. The possibly illegal large-D limit proves analytic nonidentity only; it is not the selected target.

Exactly one first-intervention coordinate is zero. Let $t_j = 0$ and choose this j as a singleton at the common ternary root. Then $D = -\ell_j$, $s_j \neq 0$, and the values $\xi_i = s_i/t_i$ for $i \neq j$ are distinct. Put

$$A_S = \log(P_S^1/P_S^0) - \ell_j.$$

Exclude $A_C = 0$ and $A_{bC} = 0$ from U. After clearing positive constants these are nontrivial affine hyperplanes in p^1, so their union has measure zero in U. In particular $A_{bC} = 0$ is equivalent to $p_j^1 = p_j^0$. For $n = 3$, $A_C = a_C t_C$ is already nonzero on the legal star region.

The common-center coordinates are all temperatures inside C, together with a_j. Equations (15)–(17) define analytic functions on the connected extension $\tau_{C \text{ vertices}} > 0$, $a_j \in \mathbb{R}$, allowing the other

slopes temporarily to be signed. A nonempty legal open subset exists near the positive-slope star after perturbing C into strict order. It suffices to prove that each required determinant is nonidentically zero on this *same* connected extension.

For designs inside C, and for b versus C in the bC refinement, set every internal C temperature to $\epsilon > 0$. Then for $i \in C$,

$$a_i t_i = A_C + \epsilon \log \frac{p_i^1 / P_C^1}{p_i^0 / P_C^0}.$$

For every internal $S \subseteq C$, the equal-temperature log-sum-exp formulas give

$$\Delta F_S^1 = A_C + O(\epsilon), \qquad \Delta F_S^2 = A_C \, \xi_{\mathrm{ext}(S)} + O(\epsilon),$$

where ext chooses the maximum ξ if $A_C > 0$ and the minimum if $A_C < 0$. The internal message divides both coordinates by ϵ. Disjoint groups have distinct extrema, giving nonzero leading coefficients of order ϵ^{-2} for internal/internal determinants and ϵ^{-1} for leaf/internal determinants. This includes b versus C, because ξ_b differs from every value inside C. Leaf/leaf determinants were already nonzero. The equal-temperature, possibly signed-slope limit is solely an analytic nonidentity witness.

The other designs are immediate. The j, b determinant is $\pm s_j t_b \neq 0$. For j versus an internal C it is $\pm s_j A_C / \tau_C \neq 0$; if C is a leaf it is $\pm s_j t_C \neq 0$. For root j versus bC, it is $\pm s_j A_{bC} \neq 0$. For root jb versus C, vary a_j so that $a_j s_j \to +\infty$ in the analytic extension. The second jb message diverges while the first C feature is fixed and nonzero, proving nonidentity. For root jC versus b, the same variation and $t_b \neq 0$ suffice.

All required determinants are therefore nonidentically zero on a connected domain. Their finite union of zero sets cannot cover its nonempty legal open subset. We have a common regular center with positive slopes and strict C edges, while the two previous responses and the adaptive final input remain fixed. This is the larger-tree step that a three-item padding argument does not supply.

Open overlap and the randomized quantifiers. The three inverse charts, restricted to the fixed first two responses, are $(n-1)$-dimensional charts in final log odds y. For refinement r, define $h_r(y) = k_{\mathrm{new}}(y) - 1$. At the center $h_r = 0$ and $dh_r \neq 0$ *on this conditional fiber*. Indeed the rank-$n-1$ slope derivative compensates a change in the new temperature while holding both previous responses fixed; equivalently use (17). The invertible full chart maps that tangent to a final-response tangent with nonzero gap derivative.

Among three nonzero linear forms on $\mathbb{R}^{n-1}$, two have a common strictly positive direction. Otherwise every pair would be oppositely proportional, impossible for three nonzero normals. A sufficiently small displacement in such a direction makes both gaps positive while preserving every other strict inequality and positive slope. An open neighborhood thus has two strict binary inverse images with identical three responses but different full functions. Identical prefixes also force identical adaptive input choices. Both conditional prior densities are positive on the overlap, so integrating their minimum gives strictly positive conditional error.

Conditional densities and null sets. In the full-rank branch, fix an admissible baseline response and hence t. Cover each topology's parameter domain by countably many submersion charts for the first-intervention map. In coordinates (z, w), z denotes the response and w the fiber. Restricting the positive conditional parameter density to a chart and conditioning on z gives a positive density in w for almost every z. Removing the countable union of exceptional sets makes this statement hold simultaneously in every chart. Thus every nonempty relatively open subset of a supported conditional fiber has positive mass. This conditional distribution and its exceptional set are fixed before the last input s is chosen. They therefore work for every subsequently selected finite s, without an uncountable intersection over inputs.

For the two regular inverse charts constructed above, change variables from fiber coordinates to final log odds y. On their common open image V, their topology-weighted conditional densities $w_r(y), w_q(y)$ are positive almost everywhere. A decoded canonical model cannot equal both distinct inverse images. Its conditional error is consequently at least

$$\int_V \min\{w_r(y), w_q(y)\} \, dy > 0.$$

This establishes positivity for the actual measurable policy; no measurable global selection of witness charts is required.

For almost every p^0, either the singular case gives positive failure on a positive-prior patch, or the full-rank case does so on a positive-probability p^1 region. Integrate over p^1, p^0, then over the previously fixed coins. Fubini yields one *fixed* target with positive failure probability. The parameter and response spaces are standard Borel, and the policy is measurable. Equality of analytic full functions can be tested on a countable dense input set. Actual error probabilities are therefore measurable; the positivity argument needs no globally selected witness pair. Countable local charts suffice. Padding handles fewer queries. This proves the theorem, not a uniform constant-error, topology-only, finite-bit, or noisy-sample bound.

Construction diagnostics, not a proof by examples. A frozen CPU check uses $n \in \{3, 4, 5, 8, 12\}$ and three first-intervention patterns: positive entries, alternating signs, and one zero. For each of these 15 cases it fixes p^0, p^1, uses a prefix-dependent final input, finds a regular common center, and constructs two strict models on an open-overlap side. Ten further checks cover untouched cherries and proportional leaf designs. All 25 pass at 110-digit precision and a separate 120-digit root-potential differentiation check. The largest forward discrepancy in the 15 constructions is below $3.16\,10^{-110}$; their extra fixed-query TV differences range from $1.75\,10^{-5}$ to $1.18\,10^{-2}$. The same research workflow performs both calculations. These are construction and implementation checks, not independent replication, calibrated error probabilities, or numerical proof of the universal quantifiers. The supplement records every prescribed center/overlap attempt and contains both calculation paths.

H Entropy proof of all-assortment stability

Fix a nonempty S and work on its pruned tree; singletons have zero TV. For a positive distribution p let P_v be its subtree masses and $q_{c|v} = P_c/P_v$. The classical nested entropy is

$$\mathcal{H}_\tau(p) = \sum_v \tau_v P_v H(q_{\cdot|v}), \qquad H(q) = -\sum_c q_c \log q_c.$$

Iterating the log-sum-exp variational identity gives

$$F_{\mathrm{root}}(u) = \max_{p \in \Delta(S)} \{\langle p, u\rangle + \mathcal{H}_\tau(p)\}.$$

Its unique interior maximizer is the model's choice distribution. The conditional entropy perspectives are concave; positive temperatures give strict concavity on simplex tangents. This representation and its learning interpretation are established prior work (Martin et al., 2022; Mertikopoulos & Sandholm, 2024).

For a tangent z, write $Z_v = \sum_{i \in v} z_i$, so $Z_{\mathrm{root}} = 0$. Differentiation gives

$$\begin{aligned} -z^\top \nabla^2 \mathcal{H}_\tau(p) z &= \sum_v \tau_v \left\{ \sum_{c:\mathrm{pa}(c)=v} \frac{Z_c^2}{P_c} - \frac{Z_v^2}{P_v} \right\} \\ &\geq t_{\min} \sum_{i \in S} \frac{z_i^2}{p_i}. \end{aligned} \tag{18}$$

Each bracket is nonnegative by Cauchy–Schwarz, and the unweighted sum telescopes. Unary brackets are zero, so menu pruning does not add a depth factor.

Subtract a common constant from $u' - u$ so its sup norm is E. Interpolate utilities and temperatures linearly by $s \in [0, 1]$. For its optimizer $p(s)$, the entropy-gradient derivative in the parameter direction, up to an irrelevant common constant, is

$$g_i = -\sum_{v \text{ on path to } i} \Delta\tau_v \log q_{\mathrm{child}(i)|v}, \qquad |g_i| \leq \delta(-\log p_i).$$

Differentiating first-order optimality and taking the inner product with $\dot{p}$ eliminates the simplex multiplier. With $V = \sum_i \dot{p}_i^2/p_i$, (18) and weighted Cauchy–Schwarz yield

$$t_{\min} V \leq \sqrt{V}\sqrt{\mathrm{Var}_p(\Delta u + g)}.$$

For $V = 0$ the claim is immediate. Otherwise, the standard deviation of Δu is at most E, and that of g is at most $\delta\sqrt{\mathbb{E}_p[(-\log p_i)^2]}$.

For any distribution on m points, its surprisal $L = -\log p_i$ obeys $\Pr(L > t) \le \min(1, me^{-t})$. Consequently

$$\mathbb{E}L^2 = \int_0^\infty 2t\ \Pr(L > t)\, dt \le (\log m)^2 + 2\log m + 2.$$

Since $\|\dot{p}\|_1 \le \sqrt{V}$, integration and division by two prove (13). The argument is valid for each S with $m = |S| \le n$, proving the simultaneous statement. Finite utilities keep optimizers interior throughout a fixed parameter path. No lower bound on individual probabilities enters the final estimate.

Width dependence with a fixed canonical root. Let a temperature-one root contain an m-way nest with temperature $t = 1/2$ and one outside item. Inside set $u_1 = t\log(m-1)$ and all other utilities zero. Send the outside utility to $-M$. The first item's conditional probability is $1/2$. Changing only the inner temperature to $t + \epsilon < 1$ makes it

$$[1 + (m-1)^{\epsilon/(t+\epsilon)}]^{-1}.$$

For fixed ϵ, as $M \to \infty$ the full-model TV converges to the conditional TV. Then as $\epsilon \downarrow 0$ its ratio to ϵ approaches $\log(m-1)/(4t)$. Thus a uniform $o(\log n)$ linear temperature bound is impossible even with the root fixed and RUM order preserved. The balanced-heavy softmax witness is standard (Nair, 2025); no sharp numerical constant is asserted.

I A COMMON REFERENCE FOR A STRUCTURAL OUTER COVER

Let $\mathcal{D}_M$ be the binary derivations represented at subset M. Distinct accepted unordered splits have distinct root partitions, giving

$$N(i) = 1, \qquad N(M) = \sum_{(A,B)} N(A)N(B).$$

Necessary clades satisfy

$$K(i) = \{\{i\}\}, \qquad K(M) = \{M\} \cup \bigcap_{(A,B)} (K(A) \cup K(B)).$$

After a root split both child derivations are selected independently in the outer-cover grammar. This establishes both recursions. Correlations in the original feasible-model set can remove grammar possibilities, but do not invalidate this overapproximation. The common clades form a laminar family because they occur in every generated tree.

Every root-reachable group occurs in some full derivation, hence cannot cross a common clade. Assign it the reference temperature of its smallest containing common clade; assign common clades their chosen positive weakly ordered temperatures. In every full derivation, noncommon edges then have equal temperatures and collapse, leaving exactly one common reference function with its chosen leaf slopes and biases. Theorem 16 applies on the unknown true derivation and this function's corresponding refinement.

For an outer parameter hull Θ and reference parameters $\bar{\theta}$, define

$$e = \max_i \sup_{\theta\in\Theta,\ x\in\mathcal{X}} |b_i - \bar{b}_i + (a_i - \bar{a}_i)x_i|, \qquad \delta = \max_M \sup_{\theta\in\Theta} |\tau_M - \bar{\tau}_M|.$$

Any safe upper bounds suffice, including interval radii and coordinatewise input bounds. A positive common temperature lower bound yields

$$\sup_{\substack{\theta \text{ compatible},\ x\in\mathcal{X} \\ \varnothing\neq S\subseteq[n]}} \mathrm{TV}(p_{\theta,S}(x), p_{\bar{\theta},S}(x)) \le \min\{1, (e + \delta C_n)/(2t_{\min})\}.$$

This compares a reference to unknown compatible models, not merely a fitted model before and after repair. It is still conditional on membership and output enclosures; a small value cannot establish either prerequisite. A common clade in the grammar need not be a strictly positive true edge if its temperature-ratio interval includes one.

J REPRODUCIBILITY AND IMPLEMENTATION DETAILS

J.1 TEACHER GENERATION AND EVALUATION

For the unknown-scale experiment, use

$$s = 910000 + 1000n + 100h + r, \quad h \in \{0, 1\}, \quad r \in \{0, 1, 2\}.$$

A permutation of $[n]$ with seed s labels a recursively balanced tree ($h = 0$) or a comb ($h = 1$). The root inverse temperature is uniform on $[0.8, 1.2]$; each internal child adds a uniform increment on $[0.2, 0.6]$. Positive baseline weights, uniform on $[0.5, 1.5]$, are normalized to probabilities, from which baseline utilities and biases are obtained by (12). Absolute slopes are 2^U with U uniform on $[-2, 2]$, using seed $s + 8000000$. The positive and alternating-sign variants share the same tree, temperatures, magnitudes and input designs. Input perturbations use seed $s + 1000000$ and evaluation inputs use $s + 3000000$. The corrected extra 13 precision-audit inputs use seed $s + 17000000$. These variants are correlated paired tests. The teacher's baseline probabilities and temperatures use seed $s + 5000000$.

A topology is correct if its nontrivial leaf clusters equal the target's clusters. Predictive error is $\mathrm{TV}(p, \widehat{p}) = \frac{1}{2} \sum_i |p_i - \widehat{p}_i|$. Temperature comparisons match clusters rather than arbitrary internal node indices. Parameter errors divide coordinatewise absolute discrepancies by $\max(1, |\theta_i|)$ after setting $k_{\mathrm{root}} = 1$. The original recovery gate uses 10^{-6} for prediction and parameter tolerances; the matched cost control uses 10^{-3}. Evaluation data are not used to choose a recovered tree.

J.2 FLOATING IMPLEMENTATION VERSUS IDEAL ALGORITHM

For each active pair, let $D = [u_A, -u_B]$ and $z = z_{AB}$. The implementation first solves its 2×2 normal equations, with a least-squares fallback for a small determinant. Pairs with nonfinite coefficients, coefficient magnitude below 10^{-10}, or nonpositive internal-child coefficients are discarded. The selection score is

$$\frac{\|D\widehat{c} - z\|_2}{\max(\|z\|_2, 10^{-14})}.$$

All remaining pairs are kept in a heap; stale entries are ignored when popped. Thus the implementation can select a false pair when numerical errors obscure a residual gap. It does not attach an exact certificate to that selection. At the end it checks the compatible temperature order with a 10^{-7} numerical tolerance. The theorem's upper bound covers unordered reduced temperatures too; the experiment deliberately uses the compatible subclass and its implementation applies that stronger final check.

The dense control uses all pairwise reciprocity equations in (19). It does not receive the slope signs or magnitudes from the teacher. Its cross-block means and hierarchy reconstruction are separate from the NCQ merge implementation. The finite-difference step is fixed before evaluation and is not tuned on the 100 prediction inputs. Both methods' time includes their own probability evaluations; the experiment contains no GPU work. Reported medians diagnose this local implementation and do not constitute a hardware-normalized performance claim.

J.3 PUBLIC COEFFICIENT FIXTURE

The Swissmetro fixture uses the published six-decimal coefficients

$$b_{\mathrm{time}} = -0.898716, \quad b_{\mathrm{cost}} = -0.856701, \quad k_{\{\mathrm{train,car}\}} = 2.053862, \quad k_{\mathrm{root}} = 1,$$

and intercepts $(-0.511953, 0, -0.167141)$ in the order train, Swissmetro, car. The three time vectors are $(1.2, 0.6, 1.0)$, $(0.8, 0.9, 1.5)$ and $(2.0, 1.2, 1.8)$. The baseline costs are $(0.7, 1.0, 0.8)$. All input vectors use that same item order. For profile index $j = 0, 1, 2$, the four-point design uses seed $1200000 + j$ and the 100 evaluation perturbations use seed $2200000 + j$. Only public fitted coefficients are used; no original subject-level data are redistributed or re-estimated.

J.4 EVIDENCE INVENTORY AND FAILURE PRESERVATION

The local companion inventory maps every table to an immutable JSON report, its protocol, and the source hashes captured at execution. The essential artifacts are G-NL3 (unknown-scale recovery), G-NL4 (three-response witnesses and independent replay), G-NL5 (matched control and public-model

fixture), G-NL6 and G-NL6R (representation precision and its redundancy-control correction), and G-NL2 (earlier calibrated statistical failure). The model count is not the sum of these overlapping audits.

During manuscript review we found that G-NL6's 13 extra inputs used the same seed stream as the teacher's baseline probabilities and temperatures. Its four original inputs did not have this dependence. We preserved the original report and reran only the 17-point control, with a newly frozen disjoint seed stream, unchanged models and unchanged accuracy criteria. Table 5 uses the original four-point results and corrected 17-point results; the counts were unchanged, but prediction errors changed. This is a disclosed correction, not an additional independent replication.

One initial G-NL5 warm-up failed because deserialized temperature-map keys were strings rather than integers. It failed before the formal runs and produced no completed result report. The execution note records this correction; the same frozen protocol was then executed with the repaired loader. An earlier exploratory JSON serialization failure is also preserved in the research record, but is not an extra successful experiment in this paper. Completed source-bound report bytes are not rewritten when the manuscript wording changes. The current materials are local research assets, not a claim of an already public repository or archival release.

J.5 Certificate protocols, arithmetic, and stopped gates

G-NL7 adds nine synthetic multiway configurations and one public ActivitySim configuration at three artificial operating points, each with double and float32 outputs. At $n = 9$ the shapes are a star, three groups of three, and a root with groups of sizes one, three, and five. Its frozen report has 54 method records; the raw failures are preserved separately from G-NL7Q's post hoc RUM repair. For the 21-item model, absolute inverse temperature is the reciprocal of the product of upstream nest coefficients. Public source/configuration files retain their BSD-3-Clause license and content hashes. The independently coded probability kernel agrees to within 9.44×10^{-16} on the checked inputs. This is a component-level adapter test, not an execution of the entire ActivitySim application or a causal policy evaluation.

G-NL9 uses the three frozen Swissmetro profiles and synthetic controls: a three-leaf strict cherry, a 10^{-8} near-collapse, an exactly flat model, a zero-perturbation design, and balanced/comb trees at $n = 4, 8, 16$. The ordinary synthetic cases have biases evenly spaced in $[-0.5, 0.5]$, slopes $(-1)^i(0.7 + 0.05i)$ for zero-based i, and internal temperature $1/(1 + 0.3d)$ at root-based depth d. Three-leaf controls instead use child inverse temperature 1.6, $1 + 10^{-8}$, or 1. Query seeds are $29091630 + n + 100h$, with $h = 0$ for balanced and 1 for comb, baseline zero, and three uniform perturbations in $[-0.15, 0.15]^n$. The three-leaf controls share seed 29091633; the zero-design case resets all perturbations to zero. These are explicit development controls, not independent random draws.

The declared enclosure for a returned positive r at relative radius η is $[r/(1 + \eta),\ r/(1 - \eta)]$, intersected with $[0, 1]$, and computed outward. The radii are 10^{-12} and 2^{-22}, not selected from prediction errors. G-NL9 permits 500 explicit partial states, 50,000 pair tests, and cooperative 12-second search checks. G-NL10 uses the same 26 transcripts plus NCQ4 outputs from all three ActivitySim profiles and profile zero of each multiway synthetic shape, in both precisions. It permits 1,024 subset states, 20,000 accepted splits, 100,000 pair tests, and cooperative 20-second search checks. Total computation gates are 180 seconds; a cap is an unresolved outcome, not evidence of incompatibility.

For G-NL11, the same six ActivitySim NCQ17 records are processed at prefixes of length four and 17; all prefixes are checked byte-for-value equal to the matching four-response records. No new teacher queries are made. The midpoint design's pseudoinverse uses a fixed relative cutoff 10^{-12}. Its binary entries are subsequently constants in outward interval products, so its numerical accuracy is not an assumed certificate. Every original row is checked after interval inversion. The state/cost limits and error radii are unchanged. All three 17-point float32 certificates exceed 0.5, failing the predeclared requirement of at least two bounds at most 0.01. That version is stopped.

Interval operations use directed Decimal contexts at 50 digits. Transcendental results are widened around correctly rounded values; square-root upper endpoints are widened as well. Binary inputs are converted exactly, and endpoint serialization uses decimal strings. This arithmetic authority follows the documented library semantics (Python Software Foundation, 2026), rather than an informal

floating tolerance. Separate 80-digit intervals check query containment and held-out model values; a high-precision recursive implementation checks selected forward values. True parameters never enter the inference client.

Risk checks use the baseline and two alternating-sign box-edge inputs, moved one binary step inward, with all nonempty menus for $n \leq 4$ and full/singleton/leave-one-out menus otherwise. The uniform quantifiers come from the theorem, not from this grid. A same-agent receipt audit checks all source hashes, reconstructs 18 small grammar cases, independently enumerates their trees and common clades, and checks that packed parameter hulls contain the unrolled branch boxes. It also verifies preservation of the earlier manuscript and protected research repositories. None of this constitutes independent expert review.

Versioned evidence. The local v0.2 claim inventory additionally links G-NL7/7Q, the global repair-only check G-NL8, the observation-to-model gate G-NL9, packed-union gate G-NL10, stopped preconditioning gate G-NL11, and their audit receipts. G-NL8's raw-to-repaired bounds must not be substituted for G-NL9–11's unknown-compatible-model bounds. Original reports and report-bound code are immutable; corrections and post hoc processing have distinct names. All these are local research assets, not an already public supplementary repository.

K NUMERICAL AUDIT

The experiments test execution of the constructive formulas and expose limitations. They are development experiments, not an independent external benchmark. All protocols and completed report-bound source files were preserved before the subsequent manuscript analysis. Teacher parameters are available only to the evaluator; each recovery client receives the numerical input vectors and a probability interface. The same 36 teachers are reused for the cost and precision audits and must not be counted as independent datasets.

K.1 UNKNOWN-SCALE RECOVERY AND A MATCHED-INFORMATION CONTROL

We use $n \in \{8, 32, 128\}$, balanced and comb trees, three seeds per size and shape, and either all-positive or mixed-sign slopes: 36 cases. Absolute slopes range from $1/4$ to 4. Inverse temperatures increase strictly down the tree. The baseline is zero and the three independent perturbations are uniform on $[-0.1, 0.1]^n$. One hundred new uniform inputs in $[-0.25, 0.25]^n$ evaluate predictions and do not select merges or tuning parameters. Appendix J gives the generator and seeds.

The floating implementation selects the legal pair with the smallest relative least-squares residual and uses stable log-sum-exp for contraction. It is an approximation to the exact equality test, not a certified implementation. All 36 cases recover the correct tree and legal parameters; their maximum predictive total variation over all evaluation inputs is 1.13×10^{-13} or less. An independent replay computes a scalar inclusive value using automatic differentiation and divides physical gradients by the recovered slopes; this provides a second numerical evaluation path, not independent peer review.

Strong classical control. A dense-Jacobian control is given the same unknown-scale probability interface, without true slopes or topology. Let H denote the Hessian in latent leaf utilities and J the Jacobian of the returned probabilities in physical inputs. Then $J_{ij} = H_{ij} a_j$. For the strict compatible family, $H_{ij} < 0$ for $i \neq j$ and $H_{ii} > 0$. Therefore

$$\operatorname{sign}(a_i) = \operatorname{sign}(J_{ii}), \quad \log|J_{ij}| - \log|J_{ji}| = \log|a_j| - \log|a_i|. \tag{19}$$

Taking negative row means of the difference matrix recovers the logarithms of relative absolute slopes in a geometric-mean gauge. Column division by these slopes followed by normalization by $p_i p_j$ gives the symmetric LCA-label matrix up to its global scale. Classical agglomeration using cross-block means recovers the hierarchy; (9) and (12) recover parameters, normalized to $k_{\mathrm{root}} = 1$. This control uses central differences with fixed step 10^{-4}, requiring the baseline and $2n$ coordinate perturbations. Both clients use one batch. This is an evaluation-point comparison, not a network-round-trip comparison.

Both methods pass all 36 matched tests: correct tree, legal parameters, maximum test TV at most 10^{-3}, and normalized slope and temperature errors at most 10^{-3}. Table 4 includes probability evaluation and recovery, using one computation thread, one warm-up per method, and three alternated

Table 4: Same teachers and probability access; 12 cases per row. Times are local CPU medians in milliseconds, not isolated-hardware service benchmarks. Both methods recover 12/12 in each row.

n	NCQ points	Dense points	NCQ ms	Dense ms	Time ratio
8	4	17	0.791	0.308	2.56
32	4	65	4.730	0.877	5.39
128	4	257	49.393	4.230	11.68

Table 5: Output-precision audit. Each branch uses the same 36 models. Seventeen responses are a redundancy control, not four-response successes. Zeros count models with at least one nonpositive returned component.

Output	Four pass	Seventeen pass	Four zeros	Seventeen zeros
float64	36/36	36/36	0	0
float32	35/36	36/36	0	0
12 decimal places	30/36	30/36	5	6
9 decimal places	29/36	30/36	6	6
6 decimal places	17/36	30/36	6	6

timing repetitions. The control is faster on this inexpensive local simulator. We make no wall-clock acceleration claim.

If each additional evaluation incurred an extra serial cost c, the measured $n = 128$ medians would cross at approximately $c = 0.179$ ms. This is only the algebraic threshold $(49.393-4.230)/(257-4)$ ms under a specified billing model. No expensive remote service was measured, and batching can invalidate this model.

K.2 Representation precision and three-response collisions

The precision audit uses the same 36 teachers, with either their original four inputs or those inputs plus 13 inputs from a separate pseudorandom stream. The five output formats are float64, a float32 round-trip, and decimal rounding to 12, 9 or 6 places, followed by row normalization. Any nonpositive returned component causes a recorded failure, without a hidden pseudocount. Success requires correct topology, legal recovered parameters and maximum TV at most 10^{-3} on the original 100 evaluation inputs. Invalid or refused models receive TV $= 1$ in aggregate prediction statistics.

The predeclared float32 four-response gate was at least 30 successes out of 36; 35 passed. Its median maximum TV, including failures as one, is 1.545×10^{-7}. The corresponding corrected 17-response value is 2.694×10^{-8}. The decimal results show that more decimal digits and relative floating precision are not interchangeable. These are finite tests of one heuristic implementation, not a worst-case precision theorem or evidence about customer-sample efficiency.

Separately, 12 constructive groups with $n = 3, 4, 6$ and four designs per size instantiate the local overlap argument in Theorem 11. Each produces two different legal strict-compatible trees whose three query responses agree to TV at most 1.943×10^{-16}. Their maximum differences over 100 other inputs range from 0.00572 to 0.10131. These witnesses are numerical existence checks, not a random-design failure-rate estimate and not a proof of the theorem's randomized quantifiers.

K.3 Public-model compatibility and a statistical failure

For a public-model check, we use the rounded fitted coefficients displayed in Biogeme's Swissmetro nested-logit example (Bierlaire, 2026). The model groups train and car and leaves Swissmetro as the other root branch. At each of three artificial time profiles, we independently implement its two-level closed-form probabilities and perturb the three cost inputs. Both NCQ and the dense control recover this interface and pass the 10^{-3} accuracy gate at all three profiles. This is *one already fitted public model at three operating points*, not three real datasets, a new behavioral fit, or validation of unobserved human counterfactuals.

Table 6: Earlier calibrated frequency experiment: 18 teachers at every level. The complete design has 17 points; only the baseline and eight fitting interventions determine the candidate. Choices are simulated through multinomial counts, not collected from people.

Choices per input	Correct trees	Legal models	Median mean TV
Exact probabilities	18/18	18/18	2.87×10^{-16}
10^4	0/18	0/18	1
10^6	2/18	3/18	1
10^8	10/18	11/18	5.18×10^{-4}

A prior calibrated experiment tested empirical multinomial frequencies, with eight fitting interventions, eight held-out interventions and a baseline. A fixed Jeffreys half-count was used there, unlike the representation audit. Table 6 reports all noise levels on its 18 separate teachers. Even with 10^8 simulated choices per input, only 10 of 18 trees are recovered, below the predeclared 12-of-18 gate. That statistical gate failed. The comparison against our older, fixed-step differential method was not a comparison against a tuned external estimator, and it cannot support state-of-the-art claims. We do not turn these frequency failures into a positive sample-complexity claim by increasing the simulation budget.

K.4 Multiway component and finite-precision certificates

ActivitySim v1.3.1 supplies a prototype-MTC hierarchy with 21 alternatives and 10 internal vertices (ActivitySim contributors, 2026). Its three- and five-way branches remove the earlier binary restriction. We execute unmodified upstream numerical function bodies with small configuration/logging adapters, preserving source hashes and license. Nest coefficients multiply along a path to give absolute temperature; separately inverting each coefficient would be wrong. Utility profiles and item-wise perturbation scales are synthetic. Tour and trip use the same nesting coefficients, so this is one external component, not two datasets or a new behavioral fit.

On three operating profiles, double-precision point recovery succeeds for NCQ4, NCQ17, and a dense 43-point control. Raw float32 success is 0/3, 1/3, and 2/3 respectively under the joint topology/legality/accuracy gate. The first two failures mainly expose an overly strict equal-edge order check. A disclosed post hoc repair enforces weak order and rebuilds biases to preserve baseline probabilities; NCQ4 and NCQ17 then pass 3/3. Their largest test TV is 3.59×10^{-7} and 5.68×10^{-8} respectively. These point results are not unknown-teacher error certificates. The dense control remains 2/3; local runtime shows no general NCQ advantage.

The separate certificate audit uses relative-error intervals of radius 10^{-12} for double responses and 2^{-22} after float32 casting and normalization. An 80-digit evaluation checks that they enclose the true responses. Inference receives no true parameters or prior parameter box. Its 50-digit directed Decimal backend (Python Software Foundation, 2026) depends on documented numerical semantics, not proof-assistant verification. Four-response local inversion intersects all certified two-row solutions. An additional verified-preconditioning variant treats a floating pseudoinverse as exact binary constants, encloses its products with the interval design and response, and trusts only the resulting interval determinant. That development variant uses four or 17 cached responses.

Table 7 reports conditional all-structure bounds on $|x_i - x_i^0| \leq 0.25$, simultaneously for all nonempty menus. Twenty-three of 26 initial single-tree-search records complete, 20 have one tree, and 19 have bounds at most 0.01. The packed variant reuses those outputs plus 12 older multiway records: 32/38 complete, including 12 structurally ambiguous rows, nine with bounds at most 0.01. These are correlated processing records, not independent tasks. Small cases are checked against exhaustive tree enumeration.

At each double-precision 21-option profile, 97 reachable subset states and 222 accepted splits represent 1,488,375 binary refinements. This is a compressed outer-cover grammar, not that many validated models. The theory includes all $2^{21} - 1$ nonempty menus; actual numerical checks use full, singleton, and leave-one-out menus at three fixed inputs, only 129 points per profile. Search costs approximately 0.66–0.70 seconds locally, without an isolated-hardware speed claim. The 17-

Table 7: Conditional prediction certificates under promised model and output-error contracts. Refusal is not zero error. Syntactic trees need not all be feasible. The 17-point algorithm uses verified preconditioning.

Case and arithmetic	Points	Complete	TV upper bound
3 options, near collapse, float32	4	1/1	$1.21\,10^{-4}$
16 options, balanced, float32	4	1/1	$9.28\,10^{-3}$
16 options, comb, float32	4	0/1	Refuse
21-option component, double	4	3/3	$\leq 4.28\,10^{-5}$
21-option component, float32	4	1/3	0.871; two refusals
21-option, preconditioned double	17	3/3	$\leq 2.17\,10^{-6}$
21-option, preconditioned float32	17	3/3	0.509–0.549

response float32 gate required two of three profiles below 0.01; none passed, and that version was stopped. Extra responses do not uniformly solve interval wrapping or deep-hierarchy conditioning.

L Proof of the local finite-precision bridge

Proposition 3 is pointwise in the fixed strictly ordered binary model and its four-input design. It does not supply uniform constants over the model class. Strengthen the generic event of Theorem 7 to require rank two for the two feature columns of every disjoint pair of true clades, and rank three for the corresponding augmented columns when they are not siblings. Any such disjoint pair can be extended to a true frontier. Disjoint nonconstant features and Lemmas 5–6 therefore establish this event outside a finite union of null sets.

Proof. Fix a design on this event. Each disjoint true-clade pair has an invertible two-row minor. For a false pair, solving any such minor leaves a nonzero remaining-row residual. Intersection of inverse boxes may reject the pair even earlier. For a true sibling pair, its exact coefficients are nonzero, and an internal-child coefficient is strictly between zero and one.

Consider the zero-width computation in increasing subset cardinality. Initially all singleton states are true clades. If every state of smaller size is a true clade, false pairs of available states are rejected by an empty inverse intersection or a nonzero residual, while true siblings create precisely their true parent. Induction thus leaves exactly $2n-1$ states and one split per internal state. No proof step follows only one greedy contraction path. Every used probability mass, chosen nonzero determinant, nonzero coefficient, and reconstructed positive temperature has a positive fixed margin. False-pair residuals or disjoint inverse values also have positive margins, of which only finitely many are needed.

On a compact neighborhood avoiding their singularities, addition, multiplication, reciprocal, logarithm, and exponential have locally Lipschitz natural interval extensions. The finite operation graph therefore produces widths $O(\epsilon)$ from observation widths at most ϵ. Every chosen inverse remains certified for small enough ϵ, and every false pair still has a contradictory interval test. True branches cannot be lost by the outer-cover theorem. Take the minimum of the finitely many admissible thresholds and induct again on subset cardinality. No spurious states appear, so the finite graph used in the continuity argument is indeed the complete computation, not a graph selected by assuming its desired result.

Top-down reconstruction uses the same smooth operations. Its canonical parameter hulls have widths $O(\epsilon)$. Midpoint parameters and weak-order clipping give a reference within $O(\epsilon)$ of the true parameters. On the fixed bounded input box, $e = O(\epsilon)$ and $\delta = O(\epsilon)$, while $t_{\min}$ remains bounded away from zero. Equation (4) yields $C\epsilon$. Each unordered disjoint pair of the $2n-1$ states is tested at most once, so the eventual regime uses at most $\binom{2n-1}{2}$ pair tests. This is not a bit-complexity claim. □

For convergent directed finite arithmetic, an accumulated rounding-enclosure width ρ gives the corresponding $O(\epsilon + \rho)$ statement when both quantities are sufficiently small and resource caps allow completion. A fixed 50-digit implementation has a rounding floor; reducing only observation error

does not eliminate it. No quantitative high-probability lower bound for random-design separation is claimed. Polytomies are excluded from this small-chart conclusion: in a flat nest every nonempty subset can be a legitimate partial aggregate, allowing $2^n - 1$ states. Appendix S extends predictive consistency without claiming a small chart. Near a collapsing edge, different reduced models can fit the same finite error bands. Neither positivity of a sibling minor alone nor a favorable predictive point fit supplies structural robustness.

M ONE-EDGE CONDITIONING AND ACTUAL AMBIGUITY

This development test fixes two eight-item hierarchies, their four inputs, and their minimum temperature 0.4. Only a selected root-to-internal edge varies, with gap $g = 1 - \tau_{\text{child}}/\tau_{\text{parent}}$ in $\{0.1, 0.01, 10^{-4}, 10^{-6}, 0\}$. At $g = 0$ the reduced model is multiway; an arbitrary binary refinement is not a unique target. The ten teachers share a query matrix; double/float32 conversion produces 20 records, not 20 independent teachers. The frozen protocol and post hoc analysis have separate reports and source identities.

Table 8: Original float32 union bounds and grammar counts. Rounded for display; 0.01 is the fixed usefulness threshold.

g	Balanced bound	Comb bound	Balanced derivations	Comb derivations
0.1	0.001842	0.047264	1	1
0.01	0.002022	0.047039	1	1
10^{-4}	0.002523	0.038416	3	2
10^{-6}	0.002310	0.037974	3	3
0	0.002320	0.037951	3	3

All 20 original runs complete with interval, parameter, and common-reference checks. The largest double-precision bound is below 1.963×10^{-7}. The balanced float32 cases all meet the usefulness gate; none of the comb cases do. At positive gaps the floating point fit has correct binary topology in 15/16 cases. The exception is the float32 comb at 10^{-6}: its topology and temperature legality fail even with small held-out predictive error.

The original fixed-bias collapsed witness fits no positive-gap transcript's full declared intervals. This failed stronger gate is preserved. In a separate post hoc audit, collapse the selected edge while holding slopes and all other temperatures fixed, and reconstruct biases to preserve baseline probabilities by (12). After making its parameters explicit floating constants, an 80-digit directed calculation verifies its four probabilities against the *original* intervals. At $g = 10^{-6}$, both original teachers and their different reduced witnesses fit these bands; both parameter sets lie in the original hulls. The unchanged balanced certificate $0.0023098514\ldots$ applies to every compatible model on the input box and all menus. The comb's $0.0379734064\ldots$ bound is not useful. Bounds below 4.537×10^{-6} and 5.076×10^{-6} comparing only the two known witness functions are different objects and are not substituted for these union bounds.

The error contract uses relative radius 2^{-22}, wider than a specified round-to-nearest cell. Consequently compatibility with these bands does not establish bit-identical IEEE outputs or impossibility under stronger numerical side information. Teacher parameters are used only in the post-inference audit, not by the learner. The witnesses are post hoc development evidence, not a rerun or success under the original failed fixed-bias protocol.

Figure 1 visualizes this fixed-input audit. SVD values are diagnostics, not certified margins.

There are 26 true frontiers for the balanced hierarchy and eight for the comb. At $g = 10^{-6}$ the smallest non-reference augmented singular values are approximately 2.5211×10^{-8} and 1.4267×10^{-8}, while the smallest true sibling two-column singular value stays near 0.023835. At exact collapse some non-reference pairs become legitimate. One contraction path can miss the separating frontier; this is why Proposition 3 treats all disjoint true clades.

N CONTROLLED PRECISION LADDER: FULL RESULTS AND PROTOCOL

Table 9 reports every budget for the eight fixed teachers. B/C denote balanced/comb, the number is leaf count; Near is the balanced eight-leaf $g = 10^{-6}$ teacher; A0–A2 are three profiles of one public ActivitySim component. All teachers and their four-input designs are reused from the preceding fixed protocols. They are not fresh independent datasets. The additional zero-design control has one distinct input repeated four times; it refuses at $b = 48$.

Table 9: Conditional TV bounds at declared relative uncertainty $\eta = 2^{-b}$. R means unresolved inverse, not zero error. Display values are rounded and do not replace outward endpoints in the evidence report.

Case	12	16	20	24	28	32	40	48
B8	R	2.33e−1	1.35e−2	8.42e−4	5.26e−5	3.29e−6	1.28e−8	5.01e−11
C8	R	1	5.49e−2	3.39e−3	2.12e−4	1.33e−5	5.18e−8	2.02e−10
B16	R	6.96e−1	3.74e−2	2.31e−3	1.45e−4	9.03e−6	3.53e−8	1.38e−10
C16	R	R	R	R	6.63e−1	3.90e−2	1.52e−4	5.92e−7
Near	R	2.23e−1	1.33e−2	8.26e−4	5.19e−5	2.78e−6	1.09e−8	4.24e−11
A0	R	R	R	1	1.62e−1	9.97e−3	3.89e−5	1.52e−7
A1	R	R	R	2.05e−1	1.26e−2	7.84e−4	3.06e−6	1.20e−8
A2	R	R	R	1	1.23e−1	7.61e−3	2.97e−5	1.16e−7

For each exact probability a 90-digit directed evaluator gives an enclosure. Let r be its midpoint and form the outward interval $[r/(1+\eta), r/(1-\eta)] \cap (0, 1]$, first checking that it contains the full evaluation enclosure. No IEEE casting, noisy frequencies, or fitted coverage radius is used. Eight teachers use 32 learning input locations, reused across budgets. Teacher evaluation and subsequent risk-audit calls are separate from the inference count.

The learner is the unchanged packed interval implementation with 50-digit arithmetic, 1,024 states, 20,000 accepted splits, and 100,000 pair tests. Per-record time is capped at four seconds, with a 300-second study cap; no cap is increased after seeing a refusal. All completed grammars must contain the teacher refinement and canonical parameters. Risk checks use baseline/full menu, one alternating box corner/full menu, and the opposite corner/leave-first-out menu, with corner coordinates one binary step inside the declared box. Independent 90-digit evaluations enclose their TV errors. Three checks are an implementation audit, not the proof of the uniform guarantee.

The 65 planned records all finish processing: 47 complete, 18 refuse, and 31 have bounds at most 0.01. There are no observed teacher exclusions or audited undercoverage events. At $b = 48$ the four ordinary strict teachers have one full topology and respectively 15, 15, 31, 31 subset states. Their approximate bounds are 5.015×10^{-11}, 2.022×10^{-10}, 1.379×10^{-10}, and 5.923×10^{-7}. The first tested useful budgets are 24, 24, 24, 40, 24, 32, 32, 32 in table order. The coarser tests do not prove necessary precision; no minimum precision or population success rate is estimated. The study takes about 26.3 local CPU seconds, a diagnostic rather than a speed claim.

Completed source-bound reports are immutable. The protocol's frontmatter used a provisional wall-clock label later than the actual report timestamp; this metadata discrepancy is recorded without changing the frozen protocol bytes. The report binds the pre-execution protocol and runner by SHA-256. Scientific outcomes, resource limits, teachers, and designs are unchanged.

O A CLASSICAL STRUCTURE FIT ON THE IDENTICAL TRANSCRIPT

The same-access dense control in Table 2 uses more input points. Here we test a different question: can a conventional numerical fit also exploit the exact same four-vector transcript? This control is an exhaustive small-tree diagnostic, not a reproduction of Aboutaleb et al. (2020)'s mixed-integer likelihood implementation or Yu et al. (2026)'s assortment learner. The latter already discusses recursive deep identification in Appendix C. Neither deep recursion nor odds cancellation alone is our novelty claim.

Fixed-tree parameterization. For each of the 15 unordered labeled binary trees on four alternatives, set

$$\tau_{\text{root}} = 1, \qquad \tau_c = \tau_{\text{pa}(c)}\sigma(z_c) \quad (c \text{ internal}), \qquad a_i = \exp(\ell_i),$$

where σ is the logistic function. The baseline probabilities determine every group mass P_c^0. At root baseline value zero, the leaf utility is

$$f_i^0 = \sum_{v \to c \text{ on the root}-i \text{ path}} \tau_v \log(P_c^0/P_v^0), \qquad b_i = f_i^0 - a_i x_i^0.$$

Thus all biases are eliminated before optimization; there are six fitted coordinates per tree. These are standard conditional-probability identities, not the unknown-tree change-message contraction. The numerical bounds are $\ell_i \in [-6, 6]$ and $z_c \in [-6, 18]$; every teacher is interior.

Using only the three nonbaseline responses, minimize

$$\frac{1}{2}\sum_{t=1}^{3}\sum_{i=1}^{3}\left[\log\frac{p_i(x^t;\theta)}{p_4(x^t;\theta)} - \log\frac{p_i^t}{p_4^t}\right]^2.$$

Each tree receives three fixed starts: unit slopes and internal child/parent temperature ratios 0.35, 0.7, and 0.97. SciPy 1.18.0 `least_squares` uses trust-region reflective steps, analytic parameter Jacobians, Jacobian-column scaling, linear loss, stopping tolerances 10^{-12} for cost, step and gradient, and at most 300 residual evaluations per start. For an internal forward value, the derivative used is

$$df_v = \sum_c w_c\, df_c + H(w)\, d\tau_v, \qquad d\log w_c = (df_c - df_v - \log w_c\, d\tau_v)/\tau_v.$$

Only the smallest training cost among the 45 fits selects the result; ties use enumeration order. The true structure and test values never enter this client or its initialization. No global-optimization certificate is asserted.

Frozen cases and outcome. Cases 0–5 use seeds $1{,}500{,}000 + 100\,\text{case}$. The teacher shapes in order are balanced, comb, balanced, comb, balanced and comb; each is randomly relabeled. The first four have inverse-temperature increments 0.6 on every internal edge. The last two replace the deepest cherry increment by 10^{-5}, leaving all others unchanged. Slopes are $\exp(\text{Unif}[-0.6, 0.6])$; baseline probabilities are the softmax of independent $N(0, 0.5^2)$ logits. Draw $x_i^0 \sim \text{Unif}[-0.15, 0.15]$ and three independent perturbations in $[-0.2, 0.2]^4$. Both methods receive identical copied input/probability arrays. After selection, evaluate 256 audit inputs in $x^0 + [-0.5, 0.5]^4$ from seed plus $1{,}000{,}000$.

Table 10: Identical four-response inputs, G-NL15. Both methods recover every true structure. Reported errors are floating evaluations, not 10^{-15} certificates. Six synthetic development cases are not six independent applications.

Case	Shape / condition	NCQ max. TV	Classical fit max. TV
0	Balanced, regular	6.80×10^{-16}	1.39×10^{-15}
1	Comb, regular	4.44×10^{-16}	3.89×10^{-16}
2	Balanced, regular	7.36×10^{-15}	5.56×10^{-15}
3	Comb, regular	6.04×10^{-16}	5.55×10^{-16}
4	Balanced, gap 10^{-5}	3.47×10^{-16}	7.63×10^{-16}
5	Comb, gap 10^{-5}	4.58×10^{-16}	4.37×10^{-16}

All 270 local optimizations terminate within the predeclared CPU budget. Both methods pass the fixed structural and TV $\leq 10^{-6}$ gate on all six cases. The new analytic Jacobians agree with centered differences to scaled maximum error below 6.67×10^{-10} on all 15 structures; baseline reconstruction and scalar-potential gradient probabilities agree within 10^{-12}. A separate conditional-probability evaluator recomputes all 270 stored objectives, selection identities and held-out predictions. This is same-workflow checking, not independent expert replication. The supplement retains every attempt and replays the complete same-transcript search.

Interpretation and limits. This is a tie, not a predictive-accuracy separation. Four exact vectors are sufficient information; they need not be useful only to NCQ. NCQ avoids exhaustive discrete-tree enumeration in its exact construction, but this experiment does not show that every classical method needs such enumeration. Its numerical point fit is not the interval certificate. Successful local termination, even on all enumerated structures, does not certify the continuous global minimum or all compatible models. The near-collapse cases still use high-precision responses, so they do not repair the float32 or frequency-learning failures. The 300-second cap is a cost guard; shared-machine timings are not reported as hardware speedups.

P Relative-temperature stability and fixed-output postprocessing

The absolute-radius bound in Appendix H allows arbitrary positive temperatures. Here an additional order assumption permits a different certificate. The tools are classical conditional-logit sensitivity, substitutability, and the entropy chain rule (Train, 2009; Martin et al., 2022; Mertikopoulos & Sandholm, 2024); we make no first-discovery claim for a general perturbation principle.

Proposition 17 (Relative-temperature stability)**.** On the same finite hierarchy, let τ and $\bar{\tau}$ be positive and weakly decrease from parent to child. Write

$$\rho = \max_v |\log(\bar{\tau}_v/\tau_v)|, \qquad t_{\mathrm{ref}} = \min_v \bar{\tau}_v, \qquad E = \min_{c\in\mathbb{R}} \max_i |\bar{u}_i - u_i - c|.$$

For every nonempty menu S, with $m = |S|$,

$$\mathrm{TV}(p_{\bar{\tau},S}(\bar{u}), p_{\tau,S}(u)) \le \min\left\{1, \frac{E}{2t_{\mathrm{ref}}} + \rho \log m\right\}. \tag{20}$$

For fixed utilities the first term vanishes. The same assertion holds for arbitrary finite arity, equal-temperature refinements, and the unary copies left by pruning.

Proof. Fix a menu and its pruned tree. Let P_v denote mass at internal node v, q_c its active child probabilities, and $H_v = -\sum_c q_c \log q_c$. Write F_v for its inclusive utility. With descendants fixed, $\partial F_v/\partial \tau_v = H_v$. Collapse v to a virtual leaf for the ancestor calculation. For an outside leaf j, differentiating the probability path product gives

$$\frac{1}{p_j}\frac{\partial p_j}{\partial F_v} = P_v \left[-\frac{1}{\tau_{\mathrm{root}}} + \sum_{\substack{w \neq \mathrm{root} \\ w \text{ common internal ancestor of } j,v}} \frac{1/\tau_{\mathrm{pa}(w)} - 1/\tau_w}{P_w} \right] \le 0.$$

Every term in the brackets is nonpositive by ordering. For nonroot v, let $c(w, v)$ be the child of strict ancestor w on the path to v. The own derivative has the nonnegative expansion

$$\frac{\partial \log P_v}{\partial F_v} = \sum_{w \text{ strict ancestor of } v} \frac{P_v/P_{c(w,v)} - P_v/P_w}{\tau_w} \le \frac{1 - P_v}{\tau_{\mathrm{pa}(v)}}.$$

The numerators telescope when temperature factors are omitted, and $\tau_w \ge \tau_{\mathrm{pa}(v)}$. Consequently $\kappa_v = \tau_v\, \partial \log P_v/\partial F_v$ lies in $[0, 1]$; at the root set $\kappa_v = 0$. For a leaf i below active child c of v,

$$\frac{\partial p_i}{\partial \tau_v} = \frac{p_i}{\tau_v}\left[-\log q_c - (1 - \kappa_v)H_v\right].$$

Outside derivatives are nonpositive because $H_v \ge 0$. The total derivative mass is zero, so its positive mass equals half its ℓ_1 norm. Since $-\log q_c \ge 0$ and $\kappa_v \le 1$,

$$\frac{1}{2}\left\| \frac{\partial p}{\partial \tau_v} \right\|_1 \le \frac{P_v H_v}{\tau_v}. \tag{21}$$

Interpolate $\tau_v(s) = \exp((1-s)\log \tau_v + s \log \bar{\tau}_v)$. Ordering persists along this path and $|\dot{\tau}_v|/\tau_v \le \rho$. The triangle inequality and the entropy chain rule on the pruned tree give

$$\frac{1}{2}\|\dot{p}(s)\|_1 \le \rho \sum_v P_v H_v = \rho H(p(s)) \le \rho \log m.$$

Integration proves the fixed-utility assertion, uniformly over finite utilities. Then change utilities at fixed reference temperatures using the utility-only part of the GS bound; additive shifts do not change probabilities. Singletons have zero TV, and unary nodes have $H_v = 0$. □

From one tree to the packed union. Keep the common-clade reference from Appendix I, with one ordered temperature per common clade and identical leaf parameters in every derivation. A non-common group inherits the temperature of its smallest containing common clade. Equal-temperature collapse therefore gives one reference function on every menu, not a separately fitted predictor for each tree. If $[\ell_M, h_M]$ is a positive hull for every reachable internal group, use

$$\rho_{\mathrm{union}} = \max_M \max\{|\log(\ell_M/\bar{\tau}_M)|, |\log(h_M/\bar{\tau}_M)|\}.$$

With the declared-domain utility radius e, a sound alternative is $B_{\mathrm{RT}} = \min\{1, e/(2t_{\mathrm{ref}}) + \rho_{\mathrm{union}} \log n\}$. Take its minimum with the original GS certificate for the same predictor. The hulls may also contain unordered assignments; RT certifies only the ordered target class already promised by the outer-cover contract. Check reference ordering explicitly. A refused search remains refused. This does not establish observation coverage, membership, or feasibility of every box assignment.

Why ordering cannot be omitted. For $((0,1),2)$, take root temperature 1, inner temperature 100, and utilities $(0, 0, 100\log 2)$. Increasing the inner temperature to 101 changes its mass from $1/2$ to $2/3$, so TV is $1/6$. But $\log(101/100)\log 3 < 0.011$. This unordered model refutes an unrestricted RT extension, not Proposition 17 or GS. An accurately known cold nest need not amplify the RT temperature term; uncertainty in utilities or in that cold nest can still dominate. There is no claim that RT always wins.

All saved precision-ladder rows (NL20). After developing the refinement, we fixed a postprocessing study on all 65 saved NL14 rows: no new queries, fits, search, parameter boxes, predictor changes, or threshold tuning. The original 47 completions and 18 refusals are retained. RT is tighter than original GS in 28 of the 47 completed rows. Combining the bounds raises the count at most 0.01 from 31 to 32 out of 65. After seeing this outcome, we declared a stronger classical control before computing it on all completed rows:

$$B_{\mathrm{splitGS}} = \min\left\{1, \frac{e}{2t_{\mathrm{ref}}} + \frac{\delta C_n}{2t_{\min}}\right\}.$$

It gives GS the same separated utility path as RT. Taking the better of original and split-path GS still gives 31/65, and RT is tighter than this stronger control in 27/47 completed rows. Table 11 identifies the sole newly useful row. No original nonvacuous bound improves by a factor of two in this study.

Table 11: One previously saved ActivitySim profile at relative error budget 2^{-28}. All methods use identical boxes and predictions. Values are TV certificate upper bounds, not measured prediction errors.

Record	Original GS	Split-path GS	RT
Profile 1, $b = 28$	0.01256236	0.01255606	0.00880722

The three profiles remain one fitted component, and the budget is not hardware precision. The unchanged predictor has not become more accurate. All adverse comparisons and refusals remain in the supplement. The modest crossing does not fix the frequency-learning failure or provide independent confirmation. An 80-digit directed calculation is checked separately at 120 digits; these arithmetic checks remain conditional on the original enclosure authority.

Constructed diagnostic and scope of numerical checks. A second fixed diagnostic uses all 15 binary five-leaf trees containing the common pair $\{0,1\}$. The reference pair temperature is 0.01, the root is 1, and noncommon groups inherit 1. Targets keep the cold pair exact and set other internal temperatures at depth d to $1 - 0.004d$. All 31 nonempty menus at three fixed utility vectors are retained: 1,395 model–menu–input checks. Some cases have original GS clipped to 1 while RT is about 0.01293; this is a constructed separation, not new learned data. The common binary refinements agree with a separately evaluated multiway reference. A scalar-potential automatic-differentiation check covers 90 model–menu–input combinations and 354 active node derivatives, with no observed violation of (21) at tolerance 10^{-10}. These same-workflow finite checks are neither an all-menu proof nor independent peer verification.

Q A CANDIDATE-ONLY SHARED-RESPONSE CERTIFICATE

The packed cover supports structural ambiguity and arbitrary finite branching, but has worst-case exponential search. Here we give a complementary, locally applicable certificate around an arbitrary proposed strict binary model. The proposal is not assumed to be the true model or to fit the observations. The mechanism is an all-competing-model exclusion using only the candidate's clades. Interval differentiation and centered range forms are classical verification tools (Rump, 2010, Sections 8 and 11); neither is claimed as a new principle.

Q.1 ONE RESPONSE BOX AND ONE FIXED GRAPH

Let θ_0 be any strictly ordered binary candidate on the labeled leaves, with root temperature one and nonzero slopes. Let p^r_{0i} be its exact response at x^r, $r = 0, 1, 2, 3$. For the supplied positive observation intervals I^r_i, compute a verified $\eta < 1$ such that

$$|p/p^r_{0i} - 1| \le \eta \quad (p \in I^r_i), \qquad \lambda = -\log(1-\eta).$$

Every compatible response table belongs to the larger family

$$p^r_i(t) = p^r_{0i} e^{t^r_i}, \qquad t \in [-\lambda, \lambda]^{4n}. \tag{22}$$

This box includes unnormalized tables; using it is a safe outer enlargement. All masses and baseline weights below are computed from the same table, with no forced normalization of its root mass. Ratios cancel row normalization.

Let $\mathcal{C}_0$ be the candidate's $2n-1$ clades. Leaves use $u_i = (x^r_i - x^0_i)^3_{r=1}$. For each candidate sibling pair A, B, put

$$D_{AB}(t) = [u_A(t), -u_B(t)], \qquad z_{AB}(t) = (\Delta \log(P_A/P_B))^3_{r=1}.$$

Choose at the center a two-row pivot with a nonzero determinant and keep those rows fixed throughout the box. The implementation chooses the largest certified central absolute determinant, with a fixed tie order. Solve the two equations by Cramer's rule, then form $u_{A\cup B}$ by (3) using the baseline mass ratios. For an arbitrary table the third equation need not hold; for a compatible model on this tree it does. Requiring every fixed pivot to be nonsingular on the whole box makes this one smooth graph, without switching inverses after perturbations.

Certified shared derivatives. Each scalar graph node stores outward enclosures of its center $f(0)$, its range on the box, and every component of its gradient on the box. Elementary chain rules propagate the gradient using the certified operand ranges. The natural range is intersected with the centered enclosure

$$f(0) + \sum_{j=1}^{4n} [\partial_j f]\,[-\lambda, \lambda]. \tag{23}$$

Both ranges enclose the same function, so their intersection is sound. The signed derivatives with respect to the same response variable can cancel before a magnitude is taken; this avoids injecting an independent worst-case error at each depth. The derivative intervals are for the whole box, not just its central Jacobian. All domains, inverses, center values and arithmetic are verified outward. A denominator containing zero or an unresolved domain returns refusal; an empty intersection is an implementation error, not evidence of incompatibility.

Proposition 18 (Shared-response graph certificate)**.** Suppose the fixed graph is verified on (22) and all of the following hold: (i) each chosen sibling pivot excludes zero; (ii) every internal-child coefficient range lies strictly inside $(0, 1)$ and every leaf coefficient range excludes zero; (iii) for every disjoint nonsibling $A, B \in \mathcal{C}_0$, the full three-row determinant

$$\det[\, u_A(t), -u_B(t), z_{AB}(t)\,]$$

excludes zero throughout the box. Then every compatible reduced weak-RUM finite-arity model has exactly the candidate's strict binary topology. Its edge coefficients lie in the verified graph ranges, which give canonical parameter boxes and the all-menu prediction bounds of Section 4 around the candidate. Acceptance does not establish that any compatible model exists.

Proof. Fix any compatible model and any binary refinement, allowing equal-temperature refinement edges. Suppose it has a clade outside $\mathcal{C}_0$, and choose a minimal such clade C. Its children A, B and all their descendants are candidate clades. A binary subtree on A has $2|A| - 1$ clades, exactly the number of candidate clades contained in A; inclusion therefore implies equality of these subtrees. The same holds for B. Within these subtrees the fixed pivots are invertible, so the model's true coefficients and normalized messages are exactly the graph's values at its shared response table. But actual siblings A, B satisfy all three coupling equations, forcing the displayed determinant to zero. They are not candidate siblings, since their union C is not a candidate clade. This contradicts (iii).

Consequently every refinement clade is a candidate clade; counting $2n - 1$ clades gives equality of the entire binary trees. Invertibility again encloses every true coefficient. Condition (ii) excludes equal-temperature refinements, proving that the original reduced model is the same strict binary tree. Starting from $\tau_{\mathrm{root}} = 1$, propagate coefficient boxes by $\tau_c = \tau_v c_c$ for internal children and $a_i = \tau_v c_i$ for leaves. Baseline conditional masses give leaf utilities through (15); actual compatible responses are normalized, so root mass is one in this parameter reconstruction. Subtract $a_i x_i^0$ for the root-baseline-zero bias gauge. These boxes enclose all compatible models simultaneously. The candidate is one fixed reference model, so the existing GS and ordered RT bounds apply on the declared input domain and every menu. No true tree or teacher parameter enters this argument. □

Local completeness and computational scope. For a fixed strict candidate and a generic four-input design, with a response tube shrinking to its exact table, the finitely many central sibling pivots, false-pair determinants and strict coefficient margins stay separated from zero. The finite graph is smooth in a neighborhood of that table; convergent outward interval ranges shrink to their central values. Thus sufficiently small tubes pass in exact-real interval arithmetic, with numerical floors in a finite implementation. The required precision is model-and-design dependent, not uniform. There are $O(n^2)$ pair checks, each with $4n$ derivative components, yielding $O(n^3)$ arithmetic work after the proposal is supplied. For the certificate alone this is not bit complexity, a guarantee of finding a good proposal, or a polynomial-time noisy learner; the next result separately specifies a locally successful ideal proposal. Genuine polytomies and retained structural ambiguity cannot pass a unique-strict-tree certificate. Failure is unresolved, not impossibility; the packed cover remains available in those regimes.

Q.2 An explicit proposal-to-certificate local bridge

The preceding certificate accepts an arbitrary proposal, but a fixed candidate's local completeness alone does not guarantee finding a suitable, observation-dependent proposal. We close this gap for an explicitly specified *ideal real-arithmetic* generator, not for the fixed-threshold experimental fitter.

Normalize each row of the positive interval midpoints to obtain y_i^r. Start with singleton groups and the leaf messages above, caching $Y_A^r = \sum_{i \in A} y_i^r$. For each active pair form $D = [u_A, -u_B]$ and $z = (\Delta \log(Y_A/Y_B))_{r=1}^3$. Discard rank-deficient D; otherwise compute

$$\widehat{c} = (D^\top D)^{-1} D^\top z, \qquad s_{AB} = \|D\widehat{c} - z\|_2^2.$$

Discard options whose internal-child coefficients are outside $(0, 1)$ or whose leaf coefficients are zero, with *no fixed positive cutoff*. Merge a minimum-score remaining pair, breaking ties by a fixed leaf-label order, using (3); set $Y_{A\cup B} = Y_A + Y_B$. If none remains, refuse. At the root recover temperatures, slopes and baseline potentials as in the exact algorithm, with root temperature one and root baseline potential zero. This gives a strict candidate $\widehat{\theta}$; apply Proposition 18 to the original intervals, using its exact candidate responses as center and a largest-absolute-determinant two-row pivot at each merge. Specifically, for this ideal procedure write $I_i^r = [L_i^r, U_i^r]$ and choose

$$\eta = \max_{i,r} \max \left\{ \left| L_i^r / \widehat{p}_i^r - 1 \right|, \left| U_i^r / \widehat{p}_i^r - 1 \right| \right\},$$

refusing if $\eta \geq 1$. This endpoint-maximum choice ensures a shrinking tube; an arbitrary valid but larger fixed radius preserves soundness, not local completeness. The candidate fits the normalized baseline row, but need not fit the other three rows.

Corollary 19 (Local proposal and verification)**.** Fix a strictly ordered binary target, a bounded prediction box and a generic four-input design as in Proposition 3. In the common root-temperature-one, root-baseline-zero gauge, there exist constants $\epsilon_0 > 0$ and $C, K < \infty$, depending on the target,

design and prediction box, such that for every positive observation box containing the true table with coordinate widths at most $\epsilon < \epsilon_0$, the ideal generator returns the true topology and canonical parameters within $K\epsilon$ of the target. The subsequent shared-response test completes, encloses all compatible models in parameter boxes of width $O(\epsilon)$, and certifies all-menu TV at most $C\epsilon$ relative to this candidate on the prediction box. The complete procedure uses $O(n^3)$ unit-cost real-arithmetic operations, not bit operations.

Proof. Write p_* for the positive true table. Validity and the width bound give $y = p_* + O(\epsilon)$, uniformly over the allowed boxes. There are finitely many true sibling-contraction histories and prefixes. Along each forced prefix, every active pair consists of disjoint true clades; on the generic event its D has rank two. Its messages and least-squares formulas are smooth near p_*. At p_* true siblings have zero residual and strictly legal coefficients, while every false pair has positive residual by augmented rank three. Take a common neighborhood over all prefixes and pairs. There, true residuals are $O(\epsilon)$ and false residuals stay bounded away from zero, so induction forces every minimum-score choice to be a true merge, regardless of ties between true pairs. Smooth downward reconstruction gives $\widehat{\theta} = \theta_* + O(\epsilon)$ in the root-baseline-zero gauge.

The candidate responses satisfy $\widehat{p} = p_* + O(\epsilon)$. By the endpoint-maximum choice and $p_* > 0$, $\eta = O(\epsilon)$ and hence the box radius in (22) obeys $\lambda = O(\epsilon)$ even though the candidate center moves. Pivot switching needs an additional argument. For a true sibling let $m > 0$ be the largest absolute two-row minor at the true center. Shrink the candidate neighborhood until each minor changes by less than $m/4$. Any maximum pivot chosen at a nearby candidate then has true-center magnitude at least $m/2$. Only finitely many full pivot selections remain, each nonsingular at the true center. At $t = 0$ their messages are the same exact candidate messages, because every candidate sibling satisfies all three equations. Thus all these graphs have a common smooth neighborhood in $(\widehat{\theta}, t)$, with bounded derivatives and denominators separated from zero. Natural interval ranges and their centered intersections have width $O(\epsilon)$ there. All fixed true-center pivots, strict coefficient margins and false-pair determinants therefore remain separated from their forbidden values; every check passes.

Proposition 18 now rules out all competing topologies and encloses *every* compatible model, without assuming that its parameters were locally close beforehand. Smooth parameter reconstruction gives $O(\epsilon)$ boxes and $O(\epsilon)$ deviations from the candidate. The bounded input box implies $e, \delta = O(\epsilon)$ in (4), while $t_{\min}$ stays positive, proving the simultaneous bound. Recomputing all active pair scores over $n - 1$ merges uses $O(n^3)$ operations with cached four-row masses; the certificate also costs $O(n^3)$. □

The constants need not be known to the algorithm and can deteriorate near degeneracy. This result excludes polytomies and assumes exact real operations, including log, exp and comparisons; a fixed arithmetic precision or time cap is not covered. In particular, the experimental proposal's fixed 10^{-10} coefficient cutoff precludes a universal local guarantee for that implementation. For $((1, 2), 3)$ with temperatures $1, 1/2$, slopes $d = 10^{-12}$, equal biases, baseline zero and perturbations e_1, e_2, e_3, the true initial coefficients are $(2d, 2d)$. Writing $L = \log((e^{2d} + 1)/2)$, with $d < L < 2d$, a false pair has fitted coefficients $(2d - L/2, d)$ and nonzero residual $L/2$. All initial pairs are filtered out despite rank-two true designs and nonzero false augmented determinants. This is a cutoff failure even with exact arithmetic, not a counterexample to the certificate or the ideal generator.

Q.3 Fixed-observation comparison and failed precursor

We reuse all 65 ladder records of Appendix N, without changing their inputs, intervals, teachers, proposal algorithm, or 0.01 usefulness threshold. These are correlated development observations, not new tasks or an independent validation set. Inference receives only the four input locations, the observed intervals, and the box $|x_i - x_i^0| \le 0.25$. The proposal is the existing floating fitter on normalized interval midpoints, with no repairs, clipping, restarts or teacher information.

An earlier candidate-clade margin certificate independently propagated norm error radii through each layer. It certified 17/65 records, 12 useful, but failed its prespecified finest-budget gate on the 16-leaf comb, where accumulated uncertainty crossed an internal coefficient's strict upper limit. That failed protocol and its complete outcome are retained; its thresholds were not changed. The

new shared-response graph is a different sufficient condition with a separately frozen gate and 100-digit directed arithmetic. It certifies 31/65 records, 27 useful, and refuses 34. All four ordinary teachers pass at both $b = 40$ and 48; the 16-leaf comb at $b = 48$ has TV bound 1.411×10^{-10}. The predeclared gate required these eight completions, that comb bound below 0.01, and at least 16 useful records; it passes.

Table 12: Same saved observations, different sufficient conditions. Each first-five row uses eight budgets; ActivitySim aggregates 24 records from one component. Counts are completed certificates / bounds at most 0.01. Packed uses the previously reported minimum of GS and RT, not only the weaker GS baseline.

Saved teacher	Packed+RT	Layerwise precursor	Shared-response
8 leaves, balanced	7/5	6/4	7/6
8 leaves, comb	7/5	4/2	7/6
16 leaves, balanced	7/5	5/4	7/6
16 leaves, comb	4/2	0/0	6/5
8 leaves, near collapse	7/5	2/2	4/4
ActivitySim profiles	15/10	0/0	0/0
Zero design	0/0	0/0	0/0
Total	47/32	17/12	31/27

Six records become useful relative to packed+RT: the three ordinary non-16-comb teachers at $b = 20$, and the 16-leaf comb at $b = 24, 28, 32$. At $b = 24$ that comb's packed search refuses, whereas the new all-model certificate is 6.776×10^{-4}. The methods are complementary, not uniformly ordered: the candidate method refuses all 24 nonbinary ActivitySim records and misses one useful near-collapse record. Selecting an available valid certificate, with its own corresponding reference predictor, would give 38 useful records in the union; taking two radii around different predictors without selecting their reference is not valid. This is certification coverage, not improved point-prediction accuracy or statistical generalization.

The main evaluation took approximately 160.39 seconds locally; this is a resource diagnostic, not a cross-run speed comparison. Every acceptance encloses the hidden teacher and passes the three fixed prediction checks. The supplement retains every refusal and both algorithms' outputs. A separate 180-digit ordinary-recursion calculation checks all 31 accepted records and 4,722 false-pair determinants. The center and two fixed shared-response perturbations give 19,944 saved-graph range checks; two fixed complex-step directions supply 13,296 derivative counterexample checks. All 186 fixed baseline perturbations are compatible and remain inside the parameter and prediction bounds. This is a different calculation path in the same research workflow, not independent-team replication. These finite checks do not prove the full-box guarantee or the all-model proposition. The mathematical guarantee remains conditional on the promised family, valid intervals, sound implementation, and successful completion.

Wrong-proposal falsification check and its coverage failure. A separately frozen 32-condition diagnostic injects candidates into the unchanged screened certificate on eight known transcripts: ordinary strict trees, a near-collapse tree, and two flat-model controls. All original intervals contain the true model under a separate 180-digit directed check. Six correct strict proposals certify; two exact flat proposals refuse the strict-order domain check. All 24 fixed wrong proposals enter the message graph and refuse: 22 at coefficient margins and two at graph-domain checks. None reaches the complete false-pair stage. Thus no wrong certification is observed, but the prespecified meaningful-falsification gate *fails*: this test does not empirically exercise the full competing-topology exclusion mechanism. We retain every receipt and do not tune or rerun the attacks to rescue that gate. These dependent, oracle-assisted checker controls are neither new learning tasks nor a proof of soundness.

R Fresh-model validation and a center-matched control

Prospective design. The 65 observations in Appendix Q were development conditions. We froze a separate protocol and source identity before executing any fresh-model test, without changing either

inference method. The core has 12 teachers: $n \in \{8, 16\}$, comb or random-split topology, and three parameter seeds. Each teacher has two new input designs and budgets $b \in \{20, 24, 40\}$, giving 24 teacher/design units and 72 correlated processing rows, not 72 independent models. An additional three random-split teachers with $n = 24$ and budgets $24, 40$ are a prespecified stress extension, excluded from the core success gate. Three flat eight-leaf teachers and a repeated-input control give 82 rows overall. No instance is screened for favorable probabilities, determinants or outcomes.

Leaf labels are permuted. Random-split trees choose each integer split uniformly; combs split off one leaf. The root temperature is one, and each internal child's inverse temperature adds an independent $U[0.15, 0.65]$ increment to its parent's inverse temperature. Slopes have random signs and magnitudes $U[0.5, 1.4]$, with biases $U[-0.4, 0.4]$. The baseline is $U[-0.2, 0.2]^n$; three independent increments are $U[-0.2, 0.2]^n$. The second design multiplies increment coordinates by fixed scales linearly spanning $[0.5, 1]$. Separate PCG64 streams fix topology, parameters, baselines and increments. Seeds, full tensors and the execution order are preserved in the supplement. The controls use separate streams, equal internal temperatures for the flat models, and four identical inputs for the zero design.

Directed 140-digit teacher evaluation generates positive probability intervals with relative budget 2^{-b}. Both inference methods receive exactly the same four inputs and intervals, not a teacher topology or its parameters. The declared prediction domain remains $|x_i - x_i^0| \leq 0.25$ and all nonempty menus. Both methods retain nominal 20-second cooperative thresholds, not equal end-to-end time budgets. Packed checks time at merge entry; post-search hulls, reference construction and GS/RT risk calculations are outside that check. The original shared-graph clock starts before proposal generation and is checked every 16 constructed scalar nodes; its final parameter/risk calculation has no clock check. Recorded call times include the post-search calculations. Neither threshold is a hard process timeout; the overall loop uses a 1,200-second cooperative window. Controls run first, followed by a fixed permutation of all core rows and then the stress rows. The paired call order alternates. No global cap is reached; local resource counts are not an isolated-hardware speed benchmark.

Outcomes and frozen decision. The core shared-response method certifies 71/72 rows, of which 66 have TV bounds at most 0.01. One coarse-budget random-split case refuses because a false-pair determinant is unresolved; five accepted coarse-budget bounds exceed 0.01. Packed+RT completes 62/72 and supplies 36 useful bounds. The 30 additional useful conditions span every one of the 12 teachers and both design families. At $b = 40$, all 24 teacher/design units are useful, with $6/6$ in each size/topology stratum. The frozen gate required at least $18/24$, at least $4/6$ per stratum, and incremental gains on at least three teachers spanning both sizes, a random-split tree and both design families. It passes without changing a cap, seed, threshold, proposal or precision.

Table 13: Fresh conditions, including every adverse control. Entries are completed certificates / bounds at most 0.01. *Centered* is the explicitly post hoc stronger packed control below, not a replacement of the frozen comparison. Each core size/topology stratum has three teachers, with two designs per teacher; repeated budgets are not independent replicates.

Stratum	Conditions	Packed+RT	Centered	Shared graph
8 leaves, comb	18	18/12	18/12	18/18
8 leaves, random split	18	18/12	18/12	17/17
16 leaves, comb	18	10/6	10/6	18/14
16 leaves, random split	18	16/6	16/6	18/17
Core total	72	62/36	62/36	71/66
24 leaves, random split	6	6/3	6/3	0/0
Flat controls	3	3/3	3/3	0/0
Repeated-input control	1	0/0	0/0	0/0

The stress extension is unfavorable: all six shared-graph runs trigger the original 20-second cooperative check, whereas packed completes all six, three useful. The threshold is not relaxed after observing these outcomes. All three genuine polytomies correctly refuse unique-strict certification; packed supplies useful common-predictor bounds for them. Neither method certifies the repeated-

input control. These findings rule out an empirical claim of uniform superiority or demonstrated scalability.

Disclosed post hoc center-matched control. The shared certificate uses baseline-centered utility deviations; the original packed bound uses separate bias/slope radii. To test this confound, after inspecting part of the running validation we froze a stronger postprocessing control. It changes neither the completed packed grammar nor its reference predictor, and does not rerun unresolved searches. For every reachable grammar edge $v \to c$, propagate, in decreasing subset-size order,

$$F_{\mathrm{root}} = [0,0], \qquad F_c \supseteq F_v + T_v \log(M_c/M_v),$$

where T_v is the saved temperature box, M_c is the observed baseline mass interval, and only M_{root} is fixed to one. Take a hull over all incoming contexts, never an intersection or one selected derivation. Along every compatible model's derivation this is its exact canonical baseline-potential recurrence; mixing contexts only enlarges the enclosure. For the unchanged reference define $h_i^0 = a_i^{\mathrm{ref}} x_i^0 + b_i^{\mathrm{ref}} - f_{\mathrm{root}}^{\mathrm{ref}}(x^0)$ and use

$$e_c = \max_i \left\{ \sup_{v \in F_i} |v - h_i^0| + 0.25 \sup_{a \in A_i} |a - a_i^{\mathrm{ref}}| \right\}.$$

The common reference shift leaves every menu distribution unchanged. Recompute GS and RT with e_c and the existing temperature boxes, and take the smaller old/new final bound for that same reference. This is a stronger sufficient bound using the existing stability results, not an additional claimed novel principle. It still yields 36 useful core conditions, so the 30 additions survive this control. It does not isolate the effects of proposal, predictor choice, search enclosures or shared differentiation from one another. The prospective gate and all original outcomes remain unchanged.

Independent calculation path and scope. All 82 observation tables are re-evaluated with ordinary 180-digit arithmetic independently of the inference and interval backend. The separate check covers 4,160 probability entries, all 71 accepted shared graphs and all 71 completed packed grammars, including parameter containment and prediction bounds. It checks 11,206 false-pair determinants, 47,748 graph values, 31,832 fixed directional derivative tests, 426 compatible baseline perturbations and additional fixed input/menu points. The centered control also receives a separate endpoint-recurrence and bound-formula audit. Finite checks are not a proof of the entire response box; the all-model guarantee is conditional on the proposition's assumptions and sound implementation. This is fresh synthetic validation within the same research workflow, not an independent team's replication, a real behavioral dataset, or a calibrated acceptance-probability estimate.

Matched enclosure ablation (post hoc). To separate enclosure effects from candidate and reference selection, we froze a one-switch ablation on all 82 known validation conditions. Both arms use the same observation-derived candidate, reference, response box, pivots, strictness checks and GS/RT calculation. The natural arm only disables each scalar node's centered mean-value intersection; it retains derivative calculations and is not a fastest-implementation timing baseline. On the common valid graph prefix, inclusion-isotonicity gives $V_{\mathrm{centered}} \subseteq V_{\mathrm{natural}}$: each elementary natural extension preserves inclusion, and intersecting with the sound centered mean-value enclosure can only tighten it. Thus, without resource or implementation failures, natural acceptance implies centered acceptance with contained parameter boxes and no larger same-reference risk bound. This is a classical enclosure comparison, not an additional novelty claim.

In the 72 core conditions, centered propagation certifies 71 with 66 useful bounds, versus 51 with 36 useful bounds for the natural arm. The 30 numerical gains span all 12 teachers and both design families, with no reverse core numerical loss. This isolates centered propagation and its tightening feedback, not signed cancellation alone. The resource tradeoff is adverse: all six 24-leaf centered runs are clock-capped, whereas the natural arm certifies four, three useful. Both refuse all four degenerate controls. All 15,343 saved common scalar ranges satisfy inclusion; 13,828 are strictly tighter. A separate 180-digit path checks accepted graphs, parameter/risk formulas and every paired summary. This reuses known synthetic conditions; it is not new independent validation or evidence of uniform superiority.

Screened graph evaluation (post hoc). The complete original AD graph has $104n - 69 + 4\sum_{g \in \mathcal{C}} |g| + 27K$ scalar nodes, where K counts the required false-sibling pairs and $\mathcal{C}$ includes all

candidate clades. Each carries $4n$ derivative components. We therefore froze one local optimization on all 82 conditions: keep the candidate construction, merged messages and parameter certificate unchanged; first evaluate each false-pair determinant by natural interval arithmetic on the certified message/change ranges. If zero is excluded, no new derivative computation is needed for that pair. Otherwise, run exactly the original AD determinant. The original AD enclosure is contained in this natural screen, so a successful screen preserves the exclusion decision; an inconclusive screen invokes the unchanged test. All required pairs must finish before certification. Successful screen records include operand provenance and range, not invented gradients. Absent resource/arithmetic failures, numerical decisions are unchanged, and common accepted parameter boxes and risk bounds are identical. This is a conservative evaluation strategy for the existing certificate, not a new interval-analysis theorem.

Both arms save the same additional clade receipts and use the same explicit cooperative 20-second clock checks. In one alternating local CPU pass, both retain the core's 71 certificates and 66 useful bounds, including the same single numerical refusal. All six 24-leaf screened runs now certify usefully in 5.45–6.34 seconds, while all six original runs remain clock-capped. These are three teachers at two budgets, not six independent models. Of 16,304 screens, 16,302 exclude zero and two fall back to AD; both fallback receipts exactly match the original computation. All four controls still refuse. The original/screened time-ratio median across the 72 jointly completed core conditions is 2.75. This descriptive, single-run timing includes common logging overhead; capped original stress times are not treated as completion times. It is neither an isolated-hardware benchmark nor a change in worst-case cubic arithmetic cost. The earlier adverse results remain intact. A separate ordinary 180-digit path checks every completed screen's endpoint recurrence and input provenance, genuine AD records, accepted parameter/risk formulas and all summaries. It is finite numerical checking, not formal whole-box verification or independent-team replication.

S LOCAL PREDICTIVE CONSISTENCY AT POLYTOMIES

The linear-state conclusion of Proposition 3 needs a strict binary tree. Predictive consistency need not. The following supporting extension concerns the complete packed algorithm, not the strict-binary shared-response certificate or a resource-bounded guarantee. It uses the same finite-coupling closure, exact finite-arity separation and common-reference bound already established above.

Allowed aggregates. Fix a reduced weak-RUM tree T with $n \geq 3$ labeled leaves and internal degrees $d_v \geq 2$. Write $L(c)$ for a child's leaf set and define

$$\mathcal{A}(T) = \{\{i\} : i \in [n]\} \cup \left\{ \bigcup_{c \in J} L(c) : v \text{ internal},\ J \subseteq \mathrm{ch}(v),\ |J| \geq 2 \right\}.$$

Each nonsingleton aggregate has a unique smallest original LCA, its *carrier* v. Assign it temperature τ_v and inclusive value $F_A = \tau_v \log \sum_{c \in J} \exp(F_c/\tau_v)$. Its message is $u_A^r = (F_A(x^r) - F_A(x^0))/\tau_v$; leaf messages remain input changes. Equal-temperature associativity makes this definition independent of a binary refinement. A compatible pair is a disjoint pair of unions of whole children of the same original vertex; either union in this pair definition may consist of a single whole child.

Proposition 20 (Fixed-polytomy predictive bridge)**.** Fix such a tree with positive temperatures and nonzero affine leaf slopes, a bounded prediction box, and a generic four-input design as specified below. There exist model-and-design-dependent $\epsilon_0 > 0$ and $C < \infty$ such that complete packed search with exact-real interval arithmetic and valid probability widths at most $\epsilon < \epsilon_0$ has precisely the subset family $\mathcal{A}(T)$ and its compatible splits. Its full derivations are exactly the binary refinements of T. Canonical parameter-hull widths are $O(\epsilon)$, and the common-reference certificate satisfies

$$\sup_{\theta \text{ compatible},\ x \in X,\ \varnothing \neq S \subseteq [n]} \mathrm{TV}(p_{\theta,S}(x), p_{\mathrm{ref},S}(x)) \leq C\epsilon.$$

This does not certify that an artificial refinement edge has exactly zero gap.

Proof. Any two disjoint members of $\mathcal{A}(T)$ can coexist in a frontier of a binary refinement. At a common carrier their child blocks are disjoint and can be refined separately. For different carriers,

an ancestor block selecting the child containing the descendant would intersect it; disjointness excludes that conflict. Complete the remaining local resolutions and contract inside each block. After flattening artificial equal-temperature edges outside active groups, the two groups share an effective parent exactly when they form a compatible pair. Otherwise their LCA v has an intervening original reduced child w on one side. An exposed group C branching there can vary while both group features stay fixed; the finite-arity false-pair argument gives

$$\partial_{F_C} \log(P_A/P_B) = P_C(k_v - k_w)/P_w \neq 0.$$

Only the adjacent original reduced gap is used, never an artificial equal edge. Nonzero leaf slopes allow these local inclusive-utility variations in the physical product input domain.

The two features depend nontrivially on disjoint coordinates and vanish at the baseline. They are linearly independent; for an incompatible pair the log-ratio feature is not their constant linear combination. Apply the analytic evaluation argument to all disjoint allowed pairs. Outside a finite union of null sets, every two-column feature design has rank two, and every incompatible augmented design has rank three. This strengthens genericity for a fixed model, not for all models simultaneously.

On this event, induction by subset cardinality at zero width creates exactly $\mathcal{A}(T)$: incompatible pairs have contradictory inverse/residual tests, whereas compatible pairs create the permitted unions. Every allowed aggregate can be recursively split compatibly. Thus every full derivation is a binary refinement of T, and every such refinement is present. All derivations at a common aggregate have the same exact message, by the carrier formula above.

Choose a nonzero two-row minor for each disjoint allowed pair. All finitely many required positive masses, absolute nonzero coefficients, temperatures and chosen minors have fixed positive margins. Incompatible residuals have nonzero margins when incompatible pairs exist; for a flat star this condition is vacuous. An artificial edge has coefficient exactly one, retained by the weak constraint $(0, 1]$. There is no division by $1 - c$ or by its zero temperature gap.

On neighborhoods avoiding the stated singularities, the natural interval extensions of the elementary operations have locally linear widths. Intersections and weak-order clipping retain the exact values. Hulls across derivations also have width $O(\epsilon)$ because their zero-width messages coincide. For sufficiently small ϵ, chosen inverses remain certified and incompatible pairs remain excluded. Taking the minimum of finitely many thresholds and inducting again in subset size prevents spurious children from appearing; the desired operation graph is not assumed in advance. The covering invariant prevents loss of any compatible split.

Top-down reconstruction along any parent path agrees at zero width. For an aggregate A with carrier v, $F_A(x^0) = F_v(x^0) + \tau_v \log(P_A(x^0)/P_v(x^0))$. If $A \subset M$ shares that carrier, the update through M telescopes to this identity. For an original whole subtree, continue inductively across its original parent edge. Equivalently, for each represented edge $M \to A$, the identity $F_A(x^0) = F_M(x^0) + \tau_M \log(P_A(x^0)/P_M(x^0))$ telescopes through equal-temperature refinements to the same canonical baseline value. Temperature and leaf-parameter images likewise coincide across paths. Their unions therefore shrink at rate $O(\epsilon)$. The completed outer-cover theorem now encloses *every* compatible model on a represented refinement, without assuming its parameters were nearby beforehand.

Every original clade is common to all refinements. A nonoriginal proper union of children can be avoided by first pairing a child inside it with one outside, so no such extra clade is common. Consequently the grammar determines the original common hierarchy without receiving it as input. Its midpoint parameters and top-down weak-order clipping are $O(\epsilon)$ from the fixed truth. Assign each noncommon group its smallest containing common clade's temperature; equal-temperature collapse gives the same reference function on every derivation. The utility and temperature radii in the common-reference bound are $O(\epsilon)$ on fixed X, and its minimum temperature stays positive. Equation (4) gives the stated all-model, all-input and all-menu bound. □

The price of structural ambiguity. Unique carrier assignment, unordered nonempty partitions and independent local binary resolutions give

$$M(T) = n + \sum_v (2^{d_v} - d_v - 1),$$
$$E(T) = \tfrac{1}{2} \sum_v (3^{d_v} - 2^{d_v+1} + 1),$$
$$D(T) = \prod_v (2d_v - 3)!!.$$

These are respectively states, compatible split entries and full syntactic derivations, not feasible parameter points or independent data samples. For each k-child subset there are $2^{k-1} - 1$ unordered splits, which gives the middle identity. Binary trees recover $M = 2n - 1$, $E = n - 1$, $D = 1$; a star has $M = 2^n - 1$ and may require exponential work. For fixed maximum arity d, $M, E = O_d(n)$ and all retained-state pairs number $O_d(n^2)$; D can remain exponential. These counts do not bound implementations enumerating all possible subset masks.

Constants deteriorate near collapse, small probabilities or singular designs. Convergent directed arithmetic gives $O(\epsilon + \rho)$, with fixed-backend rounding floors. Finite uncertainty permits nonzero refinement gaps; neither exact equality, resource completion nor feasible-model existence is certified.

T TARGETED IMPLEMENTATION DIAGNOSTICS

The following two checks were designed after external model feedback, with their protocols and scripts frozen before their first scientific execution. They reuse known records or an analytic collision; they are not additional independent learning datasets. Original algorithms, thresholds, observations and earlier failed gates are retained. No attack search, threshold relaxation, post-freeze query augmentation or resource-budget retry is used.

T.1 CHANGING UNITS WITHOUT CHANGING THE OBSERVATIONS

The fixed absolute leaf-coefficient cutoff in the experimental proposer has units. For positive coordinate scales s_i, the conjugacy

$$x'_i = s_i x_i, \qquad a'_i = a_i / s_i, \qquad b'_i = b_i$$

preserves every leaf utility and hence every probability at the corresponding input. Nevertheless, a nonzero coefficient can cross the proposer's 10^{-10} cutoff simply because the units change.

A query-only wrapper removes this artificial dependence in a controlled setting. In ideal arithmetic set $d_i = \max_r |x_i^r - x_i^0| > 0$ and $z_i = x_i / d_i$. The canonical slope is $\alpha_i = a_i d_i$ and the bias is unchanged; no baseline translation is performed. Then $d'_i = s_i d_i$, $z' = z$, and $\alpha' = \alpha$. A deterministic proposer and checker receiving these same canonical queries and observation intervals therefore have identical canonical outputs. This is classical coordinate preprocessing, not a new identifiability principle. If a coordinate never changes, the wrapper refuses rather than dividing by zero.

The implemented wrapper instead uses $d_i = 2^{\lfloor \log_2 \max_r |x_i^r - x_i^0| \rfloor}$ via binary exponent operations. For the tested $s_i = 2^{k_i}$, canonical input bytes are exactly equal and the transformations round-trip exactly, with no overflow or underflow. This strict implementation equivariance is only asserted for these dyadic units, not arbitrary positive floating rescalings.

We fix three existing eight-leaf records at budget $b = 48$: balanced, comb and near-collapse. Each has four predeclared unit representations: unchanged, all coordinates multiplied by 2^{40}, all by 2^{-40}, and the heterogeneous exponent vector $(-40, -20, 0, 20, 40, -40, 20, 40)$. Raw and wrapped arms use the same existing proposer and unchanged certificate. The raw arm proposes in the supplied units and re-expresses that candidate canonically; the wrapped arm proposes directly from canonical queries. Both then use the *same* canonical checker inputs, observation intervals and prediction box $|z_i - z_i^0| \le 0.25$. This box maps to $|x_i - x_i^0| \le 0.25 d_i$ and $|x'_i - x'^0_i| \le 0.25 s_i d_i$. Thus all arms

Table 14: Existing three models under four unit representations: 24 paired processing arms, not 12 new models. Entries count completed strict-topology certificates.

Input units	Raw proposer	Query-normalized proposer
Unchanged	3/3	3/3
All 2^{40}	0/3	3/3
All 2^{-40}	3/3	3/3
Heterogeneous dyadic	0/3	3/3
Total correlated conditions	6/12	12/12

evaluate the same physical domain, but it is not the earlier tables' uniform physical halfwidth 0.25; their numerical TV radii must not be directly compared.

All six raw refusals occur because no feasible proposal coefficients remain. For each model, all four wrapped runs have identical canonical queries, candidates and numerical certificate contents, excluding timings. All 18 acceptances pass topology, parameter-containment and three fixed prediction checks; all are useful at TV threshold 0.01 on the stated canonical box. Separate 180-digit directed evaluations verify unchanged observation coverage in all three coordinate representations; the existing prediction audit uses 130 digits. All 24 arms finish within the prespecified 240-second total and 12-second per-arm limits. This is a units-robustness development result, not a wall-clock comparison.

The wrapper does not repair intrinsically tiny dimensionless utility changes, near-collinear queries, rounding floors, or unknown output-error coverage. The fixed cutoff remains, so Corollary 19's ideal local completeness still does not apply to this finite implementation. No earlier experiment is retrospectively relabeled as using the wrapper, and no actual-float32 certification claim follows.

T.2 A COLLISION THAT REACHES THE STRUCTURAL REFUSAL STAGE

The earlier 32-condition wrong-proposal diagnostic in Appendix Q did not reach the complete nonsibling stage; that failed coverage gate is unchanged. A new, deliberately targeted test uses the explicit models $A = ((1,2),3)$ and $B = ((1,3),2)$ from Appendix F. Keep its first three inputs $0, (1,1,1), (0, \log 2, 2\log 2)$ and append $(0, 0, 10^{-4})$. The exact formulas collide on the first three inputs but differ at the fourth. For this implementation test, inputs are the recorded binary64 numbers and parameters are exact Decimal points from 240-digit directed formula enclosures; the tiny first-three numerical residuals are reported, not claimed to be exact zeros. The fourth-response TV is approximately 9.69706×10^{-6}.

Construct a B-centered log-radius 10^{-4} observation box with outward padding and a separate A-centered log-radius 10^{-12} box. A direct two-level 280-digit directed calculation and a 320-digit ordinary calculation verify that the wide box contains both models, whereas the narrow box contains A and excludes the point model B at the fourth response. The wide box therefore *does not* distinguish the trees; refusing unique certification is the correct outcome.

Inject the fixed candidates into the unchanged screened checker:

Table 15: One analytic construction, four targeted checker controls. Candidate injection is oracle-assisted, not learned recovery. Every row completes both merges and both required nonsibling checks.

Observation box	Injected candidate	Outcome
Wide, both compatible	A	Determinant-stage refusal
Wide, both compatible	B	Determinant-stage refusal
Narrow, A compatible	A	Certified, TV $\leq 1.313 \times 10^{-11}$
Narrow, point B excluded	B	Determinant-stage refusal

For each refused candidate, a compatible alternative's true cherry is a required nonsibling pair in the candidate. Its independently evaluated augmented determinant contains zero, as does the recorded checker enclosure. All three refusals occur after the complete pair loop, not at a coefficient or domain shortcut. The narrow positive control passes its parameter/prediction audit; all four finish inside the fixed 180-second total and 45-second condition limits. The targeted development gate passes without changing the old failed protocol. This checks a previously unexercised refusal path; it neither proves soundness from finite tests nor estimates a false-certification rate.

U CERTIFICATE PROFILES AND CONFIGURATION LINEAGE

U.1 EXISTING BOUNDS AT DIFFERENT ACCURACY REQUIREMENTS

Table 1 uses the threshold 0.01. To show the same evidence at other thresholds, we summarize the saved 72 core conditions without new queries, inference, models or tuning. For each declared response budget $b \in \{20, 24, 40\}$ there are the same $N_b = 24$ teacher/design units. For a method with saved bound B_j, define

$$C_b(\varepsilon) = \frac{1}{24} \sum_{j=1}^{24} \mathbf{1}\{\text{condition } j \text{ completed certification and } B_j \le \varepsilon\}.$$

Every unresolved or refused condition remains in the denominator and contributes zero at every threshold. Figure 2 is a post-hoc descriptive profile of these fixed, correlated conditions, not a population success probability or an independent validation study. The output tolerances on its horizontal axis are distinct from the input observation budgets 2^{-b}. We use exact comparisons of the saved decimal bound strings; conversion to floating point is only for drawing. The supplement records all event thresholds on $[0, 1]$, including those outside the displayed $[10^{-12}, 1]$ range, and checks the original $66/36/36/36$ totals at 0.01. At that threshold the shared counts are $18, 24, 24$ for $b = 20, 24, 40$, versus $0, 12, 24$ for each control. Across the full saved event range in this core, the shared profile is never below the other profiles; the packed and natural-only profiles do cross at some thresholds. This finite-cohort ordering is not a dominance theorem and excludes the separately reported adverse stress and polytomy conditions.

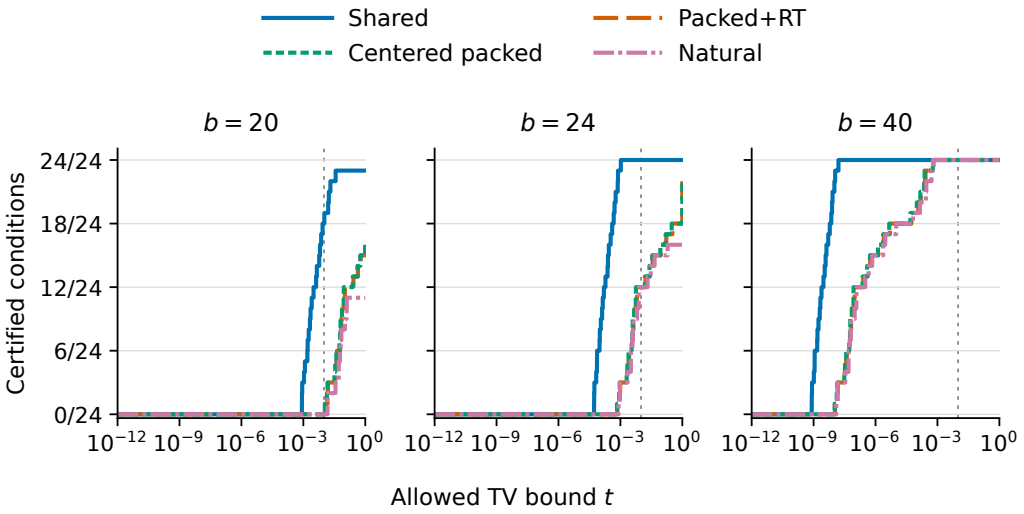


Figure 2: Certificate profiles from the existing core: 24 conditions per panel, 12 teachers overall. Shared graph and packed+RT are the frozen package comparison; centered packed and natural-only are disclosed post-hoc controls on the same observations. Failures are never removed. The vertical guide marks the original 0.01 threshold. These curves describe sufficient certificates, not point-prediction errors.

U.2 CONFIGURATIONS ACTUALLY EVALUATED

Table 16 prevents later implementation changes from being attributed retrospectively to the main validation. The core result is not a validation of one final normalized-and-screened pipeline. In the original package comparison, the shared graph uses 100-digit directed arithmetic, whereas packed search uses its existing 50-digit backend; its RT helper uses 80 digits. They share observations, prediction domain and nominal 20-second cooperative thresholds, but not the timed scope, search, reference selection or arithmetic precision (Appendix R). The stronger packed control adds 100-digit centered-utility propagation to saved grammars; it does not rerun search at higher precision. The separately frozen search-precision control below instead reruns packed search at 100 digits; it does not replace or relabel the original results. The matched natural-only ablation instead holds the 100-digit graph, candidate, reference, response tube, pivots and risk formula fixed, disabling only centered intersections while retaining gradient computations. It is this latter comparison, not the package comparison alone, that isolates the enclosure switch.

Table 16: Configuration lineage. All studies retain their own original receipts and failures. A declared response budget is not an IEEE output protocol.

Study	Evaluated configuration and scope
Fresh validation	Original midpoint floating proposer + full shared graph, versus packed+RT; no query normalization or natural screen. 72 core conditions, six stress and four controls. Original-coordinate prediction halfwidth 0.25.
Centered packed	Same 82 saved conditions, original completed grammars, boxes and reference; stronger utility postprocessing, no new search. Refusals unchanged.
Matched enclosure	Same 82 observations and same proposal in both arms; only centered intersection is disabled. Same domain and complete required checks.
Screened graph	Same 82 observations and original proposer; natural nonsibling screening with unchanged shared-AD fallback. Both arms have matched logging/clock checks. No unit wrapper.
Input units	Three old teachers, four unit representations, two arms: 24 processing arms. Raw versus query-normalized proposer; both use the screened checker in one canonical prediction box. This is not the old physical box.
Collision controls	Four oracle-assisted candidate injections into the screened checker; no learned proposal or unit wrapper. One predeclared fourth input and two boxes, not new independent models.
Search precision	Same 82 saved conditions, 50/100-digit packed search, hull/reference and GS arithmetic; RT remains at 80 digits. No shared-graph rerun or new models.

Reproduction instructions distinguish saved-result audits from executing inference. The added profile check recomputes these summaries from the packaged original rows using only the standard library. It neither re-executes the earlier experiments nor establishes the mathematical soundness of their interval arithmetic.

U.3 A PAIRED PACKED-SEARCH PRECISION CONTROL

To test the internal-precision explanation directly, a subsequent protocol fixes a paired rerun on all 82 saved validation conditions before execution. Both arms call the unchanged packed procedure on identical saved queries and observation endpoint strings, alternating 50/100-digit order. Only its directed arithmetic precision changes: this includes leaf differences, search, hull reconstruction, reference calculations and the GS bound, not just postprocessing a completed search. RT retains its separate 80-digit backend. The reference-selection rule is fixed, but reference values can shift with rounding. Neither the shared-response graph nor data generation is rerun.

The original 20-second cooperative search check and state/pair/split caps remain, with a 1,200-second total guard checked also at completion. The original post-search risk calculations are outside the cooperative search clock; this is not a hard 20-second process limit. All 164 calls and their audits finish in 46.03 seconds on the local CPU, reported as resource accounting rather than a speed comparison. The 50-digit scientific receipts match all 82 original receipts exactly after removing only elapsed-second fields. At 100 digits, no condition changes completion/refusal status, usefulness at 0.01, reachable grammar or search work counts. Table 17 retains the stress and adverse controls, with every failure in its denominator.

Table 17: Paired search-precision control on the same saved conditions. Each entry is completed / useful; useful means a conditional TV bound ≤ 0.01. This is not new-model validation.

Conditions	Count	Packed 50 digits	Packed 100 digits
Core, $b = 20$	24	16 / 0	16 / 0
Core, $b = 24$	24	22 / 12	22 / 12
Core, $b = 40$	24	24 / 24	24 / 24
Core total	72	62 / 36	62 / 36
Stress	6	6 / 3	6 / 3
Flat controls	3	3 / 3	3 / 3
Repeated-input control	1	0 / 0	0 / 0

For the 71 pairs completing in both arms, bounds do change numerically: the maximum absolute difference is 8.30×10^{-41} and maximum relative difference is 1.79×10^{-37} (rounded upward). A separate 180-digit audit path checks the 142 completed certificates, 4,160 observation values and 426 additional prediction comparisons. These finite checks are internal cross-checks, not external replication or a proof of interval soundness. Full receipts and refusal reasons are retained.

Thus the 50-versus-100-digit mismatch does not explain the observed usefulness gap on this fixed collection. This does not establish precision independence, control all package differences, or replace the matched enclosure ablation. The experiment does not test actual float32 output, alter the declared observation budgets, add independent teachers or strengthen the ideal local-completeness theorem.