FINDING IS NOT CERTIFYING: EVIDENCE REPORTS FOR REDUNDANT MECHANISMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Finding a model's minimal harmful ablation groups, ruling out additional groups under an assumption, and predicting untested interventions are different tasks. We develop evidence reports that keep these claims separate. A conditional rank frontier states which orders of additional monotone interactions have been excluded by the observations, without claiming to infer the model's true interaction order. Sharp bounds for restricted response classes show that a small record can identify an explanation under monotonicity while remaining fragile even to responses with one minimal harmful set and one recovery per intervention chain; a bound on the total number of descents changes the scaling. Fragility warnings describe compatible alternatives, not actual prediction error. We identify observed-only warnings with elementary proof completions and treat neighborhood optimization as optional rather than a default validation method. A 36-seed tiny-transformer study evaluates the actual finite-candidate routine against cheap matched controls, while overlapping-reference tests separate probability-counting, query, and optimization costs. A pretrained-model reanalysis distinguishes prompt uncertainty from mask-pair uncertainty. The resulting recommendation is to report conditional coverage, retain concrete contradiction witnesses, and use independent behavioral validation for acceptance.

1 INTRODUCTION

Removing either of two computational pathways may preserve a model's answer, while removing both changes it. Redundancy and self-repair are established phenomena (McGrath et al., 2023; Rushing & Nanda, 2024); conditional co-ablation explicitly searches for backups (Gong et al., 2026). We ask a different question: *after searching, what evidence justifies the explanation that a researcher reports?*

Three claims must be separated. A procedure can find the actual minimal harmful groups, gather enough observations to exclude additional groups under an assumption, or accurately predict the network's response to new interventions. These are not equivalent. In particular, all minimal harmful groups encode a monotone response: once an intervention contains a harmful group, the explanation predicts failure. A real network may recover after another component is removed. Even perfect group recovery then leaves response error.

A second difficulty is the assumption about interaction order. Seeing only harmful pairs does not establish that harmful triples are absent. Declaring that no interaction exceeds order two can reduce a certificate's cost exponentially, without changing the network. Instead of asking a practitioner to guess that order, we report the largest order of alternatives that the observations actually exclude. This is a conditional coverage statement, not an estimate of the network's true maximum order.

Our robustness question is equally specific. Fix the explanation and its observed labels. How bad could its untested predictions be compared with the frequency of failure followed by recovery between adjacent masks? The sharp conversion coefficient, *transcript fragility*, need not be small. Its exact full-lattice calculation is exponential and is *not* our recommended acceptance test. We separate elementary explicit-completion warnings from optional boundary-corrected optimization, and ask whether the latter changes a practical decision. Acceptance should normally use a direct, independent check of behavioral error.

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.

Table 1: Novelty boundary. We do not claim a new hypergraph model, generic flow algorithm, property tester, or scalable coverage-maximizing discovery policy.

Object	Established foundation	Addition and boundary here
Monotone completeness	Hypergraph independent covers; teaching sets (Angluin & Chen, 2006; Goldman & Kearns, 1995)	Rank/size specialization and an anytime report of excluded orders.
Fragility optimization	Binary-energy cuts and fractional programming (Kolmogorov & Zabih, 2004; Dinkelbach, 1967)	Fixed-explanation, transcript-conditioned ratio; boundary-complete partial graphs.
Robustness obstruction	Upward cube routing (Briët et al., 2012)	Sharp coefficient with one minimal failure and restricted descents; construction, not learned-model evidence.
Discovery scheduler	Harmonic set-cover charging (Chvátal, 1979)	Cross-phase certificate comparison for an exponential-time analysis baseline.

The main scientific contribution is the distinction between evidence that identifies an explanation inside a class and evidence that validates the explanation outside it. The rank frontier makes the first claim precise; sharp fragility bounds for restricted classes show the limits of the second. Classical completion, counting, and cut tools support the report but are not claimed as new scalable discovery algorithms. The empirical study tests the deployed finite-candidate routine and its cheap controls, rather than treating a larger worst-case coefficient as a useful decision by itself.

2 What an intervention record establishes

Fix a frozen network f, an input distribution P, n intervention units, and a replacement policy. For $A \subseteq [n]$, let

$$D(A) = \mathbb{E}_{x\sim P}[\ell(f(x), f^{(-A)}(x))], \qquad B(A) = \mathbf{1}\{D(A) > \tau\}, \quad \ell \in [0,1]. \tag{1}$$

A failure is disagreement with the original model above the declared threshold, not necessarily worse task accuracy. The empty mask is harmless. A query changes at most b units. Replacement policy, metric, threshold, and input distribution define the quantity being measured (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). Our units can index nodes, heads, or edges, provided every combined intervention is well defined and uses a fixed simultaneous patching rule; we do not prove such semantics for arbitrary overlapping edits. The experiments use attention heads.

Why an intervention-oracle model when weights are available? We study the evidence in a finite intervention record, rather than everything that might be deduced from the weights. A researcher may use weights or attribution to *propose* a candidate, then support a claim using recorded intervention outcomes. Our bounds describe this use of evidence, including reanalysis of a published record without rerunning the model. They do not constrain a white-box verifier that additionally proves structural properties. Such verification can certify continuous input and patching robustness (Hadad et al., 2026); it supplies stronger evidence of a different kind. Known architectural restrictions can also shrink our alternative class (Section 4). Weight access alone is not an argument that every Boolean completion is realistic, and we make no such claim.

If B is monotone, its minimal positive masks form an antichain $\mathcal{H}$ and

$$B_{\mathcal{H}}(A) = \bigvee_{e\in\mathcal{H}} \mathbf{1}\{e \subseteq A\}. \tag{2}$$

The classical hidden-hypergraph formulation applies (Angluin & Chen, 2006). For arbitrary B, all its minimal positive masks instead encode $G_B(A) = \max_{T\subseteq A} B(T)$, the least monotone upper envelope of B (Appendix H.4). Thus this representation cannot express recovery from an earlier failure.

Let $g = B_{\mathcal{K}}$ be the reported explanation and Γ its observed labels. Each $e \in \mathcal{K}$ must have an observed positive label, and its one-element deletions must be contained in observed negative masks $\mathcal{N}$. Every positive query must contain a member of $\mathcal{K}$. These are checkable conditions; they do not establish global monotonicity. A disagreement with g rejects the fixed explanation immediately.

2.1 REPORT THE RANK COVERED BY THE EVIDENCE

For declared rank r, the usual residual candidate set is

$$\mathcal{U}_r = \{T : 1 \le |T| \le r,\ g(T) = 0,\ T \not\subseteq N \text{ for every } N \in \mathcal{N}\}. \tag{3}$$

Under monotonicity and truthful labels, an empty residual identifies g among rank-at-most-r responses, provided $r \ge \operatorname{rank}(\mathcal{K})$. The bounded independent-cover characterization is given as an attributed proposition in Appendix H.

Proposition 1 (Anytime conditional rank coverage)**.** *At any record satisfying the checks above, set*

$$t_* = \min\{|T| : T \neq \varnothing,\ g(T) = 0,\ T \not\subseteq N\ \forall N \in \mathcal{N}\}, \qquad R = t_* - 1, \tag{4}$$

with $t_* = n + 1$ if no such set exists. The record identifies g in every monotone rank class $\max(1, \operatorname{rank}(\mathcal{K})) \le r \le R$. If $t_* \le n$, adding a minimizing T as a new harmful term gives a distinct consistent monotone response. For fixed verified $\mathcal{K}$, R cannot decrease as matching observations are added.

The minimum uses n binary variables, not 2^n masks: minimize $\sum_i x_i$ subject to nonemptiness, $\sum_{i \in e} x_i \le |e| - 1$ for each found edge, and $\sum_{i \notin N} x_i \ge 1$ for each negative observation. The solver either proves an optimum/infeasibility or reports unresolved optimization. Worst-case integer optimization remains hard; this is not a polynomial-time theorem. Appendix E proves the statement and specifies the report format.

Why the covered rank matters. For m disjoint harmful pairs, $\mathcal{H}_m$, and $b \ge m$, the optimal negative certificate is a binary covering array (Colbourn, 2004). In particular,

$$C_{b,2}(\mathcal{H}_m) \le 2 + 2\lceil \log_2 m \rceil, \qquad C_{b,2m}(\mathcal{H}_m) = 2^m. \tag{5}$$

The same mechanism has a cheap rank-two claim and an expensive unrestricted claim. Reporting $R = 2$ says "orders up to two have been resolved," *not* "the actual mechanism has rank at most two." An unresolved witness larger than b indicates that the intervention-size ceiling prevents the next claim; a small witness indicates a possible next query. This avoids treating a budget-limited search as complete merely because it found only small groups.

3 WHAT A FRAGILITY WARNING DOES—AND DOES NOT—SAY

Fix a monotone g and matching Γ. Let ν be a mask distribution and π a distribution on upward adjacent pairs. Define

$$e_\nu(B, g) = \Pr_\nu[B(A) \neq g(A)], \qquad \eta_\pi(B) = \Pr_\pi[B(A) = 1, B(A \cup \{i\}) = 0]. \tag{6}$$

Transcript fragility is

$$\beta_\Gamma = \sup_{B|_\Gamma = g|_\Gamma} \frac{e_\nu(B, g)}{\eta_\pi(B)}, \tag{7}$$

where $0/0 = 0$ and positive/zero is infinite. This is error against a *fixed* explanation, not distance to the closest monotone function (Goldreich et al., 2000; Khot et al., 2018). With full-support distributions it is finite exactly when Γ identifies g among all monotone responses. Rank-limited identification alone can therefore leave infinite fragility.

The exact ratio is a finite fractional binary-energy problem. Standard cut constructions compute $\max_B(e - \lambda\eta)$; fractional updates give an attaining response (Kolmogorov & Zabih, 2004; Dinkelbach, 1967). The full graph has 2^n nodes. Appendix K retains this exact small-system benchmark; the following result removes the full graph from a *warning* calculation.

3.1 BOUNDARY-COMPLETE PARTIAL GRAPHS

Choose a finite active set S containing every queried mask and the empty mask. Choose a monotone background h, for example g, constant zero, or constant one. Consider only full responses

$$B_x(A) = \begin{cases} x_A, & A \in S, \\ h(A), & A \notin S, \end{cases} \qquad x_A = g(A) \text{ on observed masks.} \tag{8}$$

This is an explicit restriction of the alternative class, not a restriction of where error is measured. In particular, every boundary edge is retained. Its violation cost is unary: an edge from an active node to an outside zero charges $x_A = 1$, while an edge from an outside one to an active node charges $x_A = 0$. Outside-to-outside edges have no descents because h is monotone.

Theorem 2 (Boundary-complete warning). *Let $E_{\text{out}} = \Pr_\nu[A \notin S, h(A) \neq g(A)]$. Maximizing $e_\nu(B_x, g)/\eta_\pi(B_x)$ over (8) gives a lower bound $\underline{\beta}_{S,h} \leq \beta_\Gamma$. For known E_{out} and evaluable mask/edge weights, each fractional-programming subproblem is a minimum cut with $|S| + 2$ vertices and $O(n|S|)$ arcs. It returns a full-domain witness. Replacing E_{out} by a certified lower bound preserves a valid lower bound on β_Γ. Enlarging S cannot weaken the exact subclass optimum.*

The proof is the decomposition of the *full* numerator and denominator into inside and boundary terms, followed by the standard cut reduction (Appendix F). A finite fractional search terminates or abstains at a declared cap; no strongly polynomial bound on its number of cuts is claimed.

For $h = g$, outside error is zero. For constant backgrounds, the only global quantity is the positive mass of the fixed reference. Disjoint groups admit a closed form. For overlapping groups we can use conservative union bounds or classical importance sampling (Karp et al., 1989); a single probability interval supports every active set subsequently chosen for that same reference and mask distribution (Appendix B). The interval concerns counting the *reference*, not measuring the network. Product mask measures use ablation probability p per coordinate; adjacent pairs choose their added coordinate uniformly and draw other coordinates independently with probability p.

Proposition 3 (Observed-only warnings are proof completions). *Let P_Γ and N_Γ be the positive and negative observed masks, with $\varnothing \in N_\Gamma$. For $S_0 = P_\Gamma \cup N_\Gamma$, the constant-background responses are exactly*

$$B_-(A) = \mathbf{1}\{A \in P_\Gamma\}, \qquad B_+(A) = \mathbf{1}\{A \notin N_\Gamma\}. \tag{9}$$

These are the two completions used in the unrestricted lower-bound proof. Their disagreement supports are respectively the unqueried masks that g predicts positive and negative. Their exact errors are $\Pr_\nu[g = 1] - \nu(P_\Gamma)$ and $\Pr_\nu[g = 0] - \nu(N_\Gamma)$; their descent masses are computed from the boundaries of the observed sets. No labels are optimized and no cut is needed.

Thus S_0 is an explicit baseline, not evidence that a scalable optimizer discovered additional fragility. Adding free neighbors gives the optional S_1 optimization. Its benefit must be compared with explicit completions and simple targeting rules on the same record. Appendix A proves the identity and a strict separation that is itself attained by a cheap closed-form completion.

A natural active set is the observed masks, optionally plus their Hamming-distance-one neighbors; its size is at most $(q + 1)(n + 1)$. We do *not* close it under arbitrary unions and intersections. Taking that closure can recreate an exponential lattice.

Why simply deleting the boundary is invalid. With four units, $g \equiv 0$, and the full mask observed harmless, the exact coefficient is seven. Keep the observed masks and their one-step neighbors: empty mask, singletons, triples, and full mask. Label the singletons positive and the other retained masks negative. There is no retained adjacent descent, so an induced-subgraph ratio is infinite. Every full extension must account for the missing boundary, and the full optimum is only seven. An extendible labeling alone therefore does not establish a lower bound on the full ratio. Our construction keeps the terms this shortcut discards.

3.2 Warning, falsification, and acceptance are different outputs

A large $\underline{\beta}$ establishes that the record permits a fragile completion. It neither proves that the actual model realizes that completion nor supplies an upper bound on actual error. In particular, substituting $\underline{\beta}$ into $e \leq \beta\eta$ would be invalid. Our report labels it a *warning*, not a behavioral certificate.

For falsification, choose an unqueried mask on which the returned completion differs from g. Observe its actual label. A contradiction rejects the fixed explanation and ends that run; a match eliminates that completion, after which a new witness can be computed. Selection uses only g and the accumulated record, never the true unqueried response. At large n, a concrete candidate can be searched within the returned finite active/boundary representation or sampled from its explicit background; the

experiments separately evaluate full-support enumeration and the actual finite-candidate search, together with a matched finite-domain control.

For acceptance, independently sample masks from the intended ν and directly bound $e_\nu(B, g)$ with a binomial confidence interval (CI) (Clopper & Pearson, 1934). Sampling adjacent pairs independently also estimates η, but neither estimate requires computing β. We recommend the direct error check for approximate acceptance. A fresh holdout must follow freezing the explanation; repeated adaptive reuse is not justified (Dwork et al., 2015; Howard et al., 2021). Appendix E gives a simultaneous error-allocation option for multiple report times.

4 Which restrictions actually reduce fragility?

Unrestricted completions are deliberately adversarial. Under uniform measures a q-query record satisfies $\beta_\Gamma \geq (2^n - q - 1)/(2q)$, and the all-harmless explanation with one harmless full-mask query has exact coefficient $2^{n-1} - 1$ (Appendix L). This means that identification under an assumption can be cheap while robustness without it is not. It is not a lower bound for direct approximate validation or white-box verification.

The next result asks whether the obstruction disappears for much simpler responses. A *chain descent* is a $1 \to 0$ transition along one sequence that adds each component once. Bounding descents per chain is different from bounding the total number of descending lattice edges, a distinction related to inversion complexity (Morizumi, 2008).

Theorem 4 (One minimal failure, one recovery per chain)**.** *Let $n \geq 2$, $g \equiv 0$, and $\Gamma = \{([n], 0)\}$, with the harmless empty mask known. Restrict alternatives to responses with at most one inclusion-minimal positive set. Their exact fragility is*

$$\beta_\Gamma^{\text{one}} = \frac{n(2^{n-1} - 1)}{2(n-1)}. \tag{10}$$

The value is unchanged if alternatives must additionally have a singleton minimal positive set and at most one descent on each increasing chain. It is attained by $B_i(A) = \mathbf{1}\{i \in A,\ A \neq [n]\}$.

The attainer has $2^{n-1} - 1$ positive masks and $n - 1$ descending edges. For the upper bound, a response with unique minimal positive T can only be positive on the $(n - |T|)$-dimensional face above T. Upward-path flow bounds the number of positives per descending edge by $(2^{n-|T|} - 1)/(n - |T|)$; the largest face occurs for a singleton T. Appendix G provides the full proof.

The attaining response is a constructed intervention truth function, not a learned mechanism. Its compact threshold/rectified linear unit (ReLU) representation uses input-independent hidden values and is therefore only a circuit-description result (Appendix D). Here "rank one" refers to the minimal-positive family, not to a monotone rank-one disjunctive normal form (DNF) representation.

A stronger descent restriction helps. Within the same one-minimal-set class, allowing at most $k \geq 1$ descending edges *in the entire lattice* gives the exact coefficient

$$\beta_\Gamma^{\text{one, total}\leq\text{k}} = \frac{n(2^d - 1)}{2d}, \qquad d = \min(k, n - 1). \tag{11}$$

For fixed k this is linear, not exponential, in n. For $k = 1$ it is $n/2$. The restriction is far stronger than one recovery per chain: the construction in Theorem 4 has $n - 1$ total descending edges. This distinction prevents interpreting every notion of "few nonmonotonicities" as equivalent.

Known effective dimension helps too. Suppose every admissible response and g depend on the same *known* j coordinates. Such functions are juntas (Blais, 2009); we do not infer the relevant coordinates from the transcript. Projection gives $e_n = e_j$ and $\eta_n = (j/n)\eta_j$, hence

$$\beta_{\Gamma,n}^J = (n/j)\beta_{\Gamma|_J,j}. \tag{12}$$

At $n = 30, j = 4$ the one-query value is 52.5, compared with $536{,}870{,}911$ without that restriction. This is an example of architectural information that materially changes the answer. The same appendix gives a multi-query restricted construction and shows explicitly how nonuniform weights affect it.

Table 2: Detection within eight additional labels among initially wrong checked explanations. Primary inference uses the 30 training-seed clusters 6–35; the final column is descriptive over all 36 clusters. Intervals resample whole seed clusters, retaining related models, patch settings, discovery methods, and repetitions together. Values are percentages; intervals are pointwise. No claim of equivalence follows from an interval containing zero.

Policy	Primary seeds: rate [95% CI]	All seeds
Uniform	48.9 [42.4, 55.5]	47.0
Predicted positive	56.9 [49.4, 64.2]	56.3
Predicted negative	32.2 [27.2, 37.2]	32.1
Cheap finite positive	57.9 [50.2, 65.5]	57.1
Explicit S_0 support	47.9 [41.5, 54.6]	47.1
S_1 full support	53.9 [47.9, 60.2]	53.2
S_1 finite candidates	54.2 [47.7, 60.8]	53.8

5 Experiments: evidence, decisions, and costs

We separate exact mathematical checks from discovery, falsification, and behavioral validation. All mask-selection policies use only the accumulated labels and the fixed explanation. Unqueried neural labels are hidden from the selector. The tiny models retain the same constructed task to isolate the change in seed count and policy implementation; this replication does not establish prevalence across tasks.

5.1 Seed-cluster evaluation of the finite-candidate routine

The study uses 36 training seeds, yielding 72 two-layer, eight-head transformers, 144 zero/mean-patch conditions, and 720 initial evaluation records. Seeds 0–5 reuse previously trained models; primary inference uses the independently trained seeds 6–35, and the 36-seed aggregate is descriptive. Each training seed covers two head-dropout settings, two patch policies, and five full-recovery procedures. We retain all models, including poorly performing ones. Full response tables provide exact evaluation on the same 64 held-out sequences per model. They are available to the evaluator, not the query policy.

For every checked record, including correct explanations, we run seven policies with four paired random repetitions and an eight-label horizon. A contradiction ends support for that fixed explanation immediately. The policies include uniform, predicted-positive, predicted-negative, the explicit S_0 completion, full-support S_1, and the *actual finite-candidate* S_1 implementation. A matched cheap finite rule uses the same candidate domain but filters only by the explanation's positive prediction. This isolates both selection and candidate-domain effects. Shorter budgets are within-run prefixes, not independent experiments.

The primary seed-cluster difference (S_1 finite minus cheap finite positive) is -3.72 percentage points with a 95% seed-cluster interval [-9.97, +1.95] and a paired Monte Carlo sign-flip $p = 0.24$. Against unrestricted predicted-positive sampling, the difference is -2.72 points [-8.76, +2.78]. The primary contrast compares finite S_1 with the cheap finite-positive control on seeds 6–35; other contrasts are descriptive. No preregistration is claimed. These data do not justify the cut optimizer as the default falsification policy. We recommend the cheap rule, not a more expensive policy merely because its worst-case warning is larger.

Charged portfolio computation averages 0.71 seconds per S_1 finite run, plus 1.34 milliseconds for candidate search; the matched cheap finite rule uses 0.58 milliseconds for candidate construction and no cuts. Mean additional labels are 6.08 and 5.87, respectively, across all checked records. These are implementation-specific costs, not an optimized-runtime lower bound.

The finite routine searches the active nodes, their immediate boundary, and reported minimal groups; it does not enumerate the cube. On eight units this domain often includes much of the cube. Support overlap, fallback rate, and candidate-search time are therefore relevant alongside detection rate. Cut costs are measured on previously unseen record states and charged to every trajectory that would

Table 3: Overlap stress test. Each mass method has 36 paired S_0/S_1 comparisons. "Zero" counts a vacuous warning, not low true fragility; infinity is reported separately. These tests do not establish a gain in neural falsification.

Mass calculation	$S_1 > S_0$	Zero (S_0/S_1)	S_1 median s	Max s
Union bounds	27/36	0/0	2.12	13.36
Reference interval	35/36	0/0	2.27	11.92

need that state; cached wall times are not presented as cold-run costs. Cheap policies receive the same extra-label budget, without a cut allowance.

5.2 Overlapping references and records that change the warning

The stress test uses overlapping cyclic, hub, and random triple references at 20, 26, and 30 units, two mask densities, and two record types. Minimal records contain positive groups, their negative one-element deletions, and a maximal harmless observation. Boundary-enriched records additionally contain all size-$(n-2)$ masks. These are constructed controls, not neural discoveries. Across 144 combinations we compare conservative union bounds with a reference-only probability interval, using the *same* observed record for both active sets. Every returned warning retains all boundary terms.

For the 26-unit random-triple reference under uniform masks, the interval-supported warning grows from 1,211,442.3 to 3,251,340.4 (2.68-fold). These ordinary stress records need not identify the unrestricted monotone class, so a finite lower bound is not a finite estimate of exact fragility. At low mask density, a large ratio can instead be driven by very rare high-cardinality descent edges; it is not a measured neural error. The reference interval removes vacuity in individual constant-background bounds in the tested cases, but the union-bound portfolio is also nonzero. The fully identifying overlapping control described in the main text also admits a cheap upper-tail completion.

A fully identifying 30-unit overlapping-block control is more informative about scope. With a complete negative cover and boundary-enriched observations, S_1 grows the warning from 11.16 million to 28.50 million, but a closed-form upper-tail completion attains exactly the same value. The cut calculation takes 35.6 seconds in the recorded run. Without the added boundary observations, S_0 and S_1 coincide. Thus even this favorable control does not establish that an optimizer is necessary.

There is also a transparent favorable case: for $g \equiv 0$, observe all pairs and the full mask as harmless. Let $H = 2^n - \binom{n}{2} - 2$. Then $S_0 = H/(2n)$, while a completion available in S_1 attains $(H-n)/2$, an order-n improvement. At 30 units the measured values are $17{,}895{,}689.8$ and $536{,}870{,}678.5$. However, the explicit rule $B(A) = \mathbf{1}\{2 < |A| < n\}$ achieves the latter without a cut. This establishes that expanding the completion class can help, *not* that optimization is needed (Appendix A).

5.3 Rank-report computation on overlapping groups

We test the mixed-integer report on 27 overlapping-reference records with 30, 60, and 100 units, using 32, 128, or 256 additional maximal harmless draws. The time limit is 0.25 solver seconds, while reported wall time also includes setup. Only a proved optimum or infeasibility releases a rank; a timeout remains unresolved. The solver finishes 9/27 instances and reports unresolved optimization on 18/27. Median wall time is 0.28 seconds and the maximum is 0.38 seconds in this execution. At 100 units only one of nine instances finishes within the declared limit. Raising the limit to five seconds solves 8/9 of those 100-unit cases. Thus avoiding a 2^n table does not make worst-case rank certification easy. The returned mask and all claimed exclusion checks are independently checked; solver optimality is still a numerical solver result, not a separately formalized proof.

5.4 Pretrained-model evidence

The Generative Pre-trained Transformer 2 (GPT-2) small-model measurement uses the published 26-head indirect object identification (IOI) set (Wang et al., 2023), 32 fixed prompts, and zero ablation. Other heads and multilayer perceptrons (MLPs) remain intact. It is not a mean-ablation reproduction of the published circuit or an evaluation over the IOI task population. At $\tau = 0.10$ and uniform

adjacent-pair sampling, two of 256 pairs show recovery; 218 smaller masks are harmful, giving a conditional recovery rate of $2/218 = 0.92\%$ (fixed-prompt pair interval 0.11–3.27%).

We resample prompts jointly across all masks, stratifying within the four templates, and report the added head for every recovery. The two primary added heads are $(2, 2)$ and $(10, 6)$; discrepancy changes are $4/32 \to 2/32$ and $7/32 \to 3/32$. Each event's paired prompt-bootstrap interval for the decrease includes zero. The occurrence claim therefore concerns only the exact finite prompt population. Template-stratified resampling does not supply new templates or hundreds of independent prompts. The full observed record contains monotonicity contradictions, so its evidence-only report rejects the monotone class rather than reporting a rank or warning for a nonexistent consistent explanation.

The pretrained analysis is limited to 32 fixed prompts and zero ablation; mean replacement and independent validation of end-to-end pretrained discovery were not evaluated. Appendix C details the uncertainty analyses. Empirical support for the complete workflow comes from the small-model study.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and polish writing. Additionally, we used generative AI tools for editing code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Hasan Abasi, Nader H. Bshouty, and Hanna Mazzawi. On exact learning monotone DNF from membership queries. arXiv:1405.0792, 2014. URL https://arxiv.org/abs/1405.0792.

Matthew Aldridge, Oliver Johnson, and Jonathan Scarlett. Group testing: An information theory perspective. *Foundations and Trends in Communications and Information Theory*, 23(1–2): 1–221, 2026. doi: 10.1108/FTCIT-11-2025-0150. URL https://doi.org/10.1108/FTCIT-11-2025-0150.

Dana Angluin and Jiang Chen. Learning a hidden hypergraph. *Journal of Machine Learning Research*, 7(79):2215–2236, 2006. URL https://jmlr.org/papers/v7/angluin06a.html.

Eric Balkanski, Oussama Hanguir, and Shatian Wang. Learning low degree hypergraphs. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 419–420, 2022. URL https://proceedings.mlr.press/v178/balkanski22a.html.

Eric Blais. Testing juntas nearly optimally. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pp. 151–158, 2009. doi: 10.1145/1536414.1536437. URL https://doi.org/10.1145/1536414.1536437.

Jop Briët, Sourav Chakraborty, David García-Soriano, and Arie Matsliah. Monotonicity testing and shortest-path routing on the cube. *Combinatorica*, 32(1):35–53, 2012. doi: 10.1007/s00493-012-2765-1. URL https://doi.org/10.1007/s00493-012-2765-1.

Ting Chen. More efficient algorithms for searching for several edges in a hypergraph. *Indian Journal of Pure and Applied Mathematics*, 57(2):425–433, 2026. doi: 10.1007/s13226-024-00561-z. URL https://doi.org/10.1007/s13226-024-00561-z.

V. Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4 (3):233–235, 1979. doi: 10.1287/moor.4.3.233. URL https://doi.org/10.1287/moor.4.3.233.

Laurence A. Clarfeld and Margaret J. Eppstein. Group-testing on hypergraphs with variable-cost tests: A power systems case study. arXiv:1909.04513, 2019. URL https://arxiv.org/abs/1909.04513.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Charles J. Colbourn. Combinatorial aspects of covering arrays. *Le Matematiche*, 59(1–2):125–172, 2004. URL https://lematematiche.dmi.unict.it/index.php/lematematiche/article/view/166.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Werner Dinkelbach. On nonlinear fractional programming. *Management Science*, 13(7):492–498, 1967. doi: 10.1287/mnsc.13.7.492. URL https://doi.org/10.1287/mnsc.13.7.492.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. In *ACM Symposium on Theory of Computing*, 2015. URL https://arxiv.org/abs/1411.2664.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Yigit Efe Erginbas, Justin Singh Kang, Elizabeth Polito, and Kannan Ramchandran. Adaptive sparse möbius transforms for learning polynomials. arXiv:2602.06246, 2026. URL https://arxiv.org/abs/2602.06246.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. arXiv:2301.04709v4, 2025. URL https://arxiv.org/abs/2301.04709v4.

Sally A. Goldman and Michael J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995. doi: 10.1006/jcss.1995.1003. URL https://doi.org/10.1006/jcss.1995.1003.

Oded Goldreich, Shafi Goldwasser, Eric Lehman, Dana Ron, and Alex Samorodnitsky. Testing monotonicity. *Combinatorica*, 20(3):301–337, 2000. doi: 10.1007/s004930070011. URL https://doi.org/10.1007/s004930070011.

Zhiren Gong, He Lu, Tiantong Wang, Yichi Zhang, Yixin Wang, Zihao Zeng, Ming Xiao, Chau Yuen, and Wei Yang Bryan Lim. Conditional co-ablation: Recovering self-repair backups in transformer circuits. arXiv:2607.01940v3, 2026. URL https://arxiv.org/abs/2607.01940v3.

Itamar Hadad, Guy Katz, and Shahaf Bassan. Formal mechanistic interpretability: Automated circuit discovery with provable guarantees. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.16823.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2403.17806.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. arXiv:2404.15255, 2024. URL https://arxiv.org/abs/2404.15255.

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

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

Justin S. Kang, Yigit E. Erginbas, Landon Butler, Ramtin Pedarsani, and Kannan Ramchandran. Learning to understand: Identifying interactions via the möbius transform. arXiv:2402.02631, 2024. URL https://arxiv.org/abs/2402.02631.

Richard M. Karp, Michael Luby, and Neal Madras. Monte-carlo approximation algorithms for enumeration problems. *Journal of Algorithms*, 10(3):429–448, 1989. doi: 10.1016/0196-6774(89)90038-2. URL https://doi.org/10.1016/0196-6774(89)90038-2.

Subhash Khot, Dor Minzer, and Muli Safra. On monotonicity testing and boolean isoperimetric-type theorems. *SIAM Journal on Computing*, 47(6):2238–2276, 2018. doi: 10.1137/16M1065872. URL https://doi.org/10.1137/16M1065872.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004. doi: 10.1109/TPAMI.2004.1262177. URL https://doi.org/10.1109/TPAMI.2004.1262177.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. arXiv:2403.00745, 2024. URL https://arxiv.org/abs/2403.00745.

David Lindner, János Kramár, Sebastian Farquhar, Matthew Rahtz, Thomas McGrath, and Vladimir Mikulik. Tracr: Compiled transformers as a laboratory for interpretability. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2301.05062.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. arXiv:2307.15771, 2023. URL https://arxiv.org/abs/2307.15771.

Hiroki Morizumi. A note on the inversion complexity of boolean functions in boolean formulas. arXiv:0811.0699, 2008. URL https://arxiv.org/abs/0811.0699.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in/.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42836–42855, 2024. URL https://proceedings.mlr.press/v235/rushing24a.html.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M. Blei. Hypothesis testing the circuit hypothesis in LLMs. arXiv:2410.13032, 2024. URL https://arxiv.org/abs/2410.13032.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL https://jmlr.org/papers/v15/srivastava14a.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.03762.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.00593.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16042.

A Explicit baselines and what optimization can add

A.1 Proof of Proposition 3

Every node of S_0 has its observed label. With background zero, the only positive nodes are the observed positives P_Γ, so the response is exactly B_-. With background one, the only negative nodes are the observed negatives N_Γ, so the response is exactly B_+. The reference background returns g everywhere and has zero error. Matching observations imply $P_\Gamma \subseteq \{g = 1\}$ and $N_\Gamma \subseteq \{g = 0\}$. Therefore the two disagreement sets are

$$\{g = 1\} \setminus P_\Gamma, \qquad \{g = 0\} \setminus N_\Gamma.$$

Their masses are the expressions in the proposition. A descending edge of B_- leaves P_Γ for its complement; a descending edge of B_+ enters N_Γ from its complement. At most $n|\Gamma \cup \{\varnothing\}|$ incident edges need examination. Once the reference's positive mass is supplied, these are direct evaluations, not cut problems. With an approximate mass lower bound the reported number is conservative, but the returned full responses are still the same explicit completions.

This identifies the observed-only routine with the two responses already used in Appendix L. Under uniform masks, their error counts sum to $2^n - q - 1$ and their total descent count is at most nq, yielding the familiar lower bound $(2^n - q - 1)/(2q)$ when $q \geq 1$. Recomputing the same response with a small graph is not a separate algorithmic contribution. Likewise, uniform sampling from its disagreement support is exactly predicted-positive or predicted-negative targeting, except for empty-support fallback. It is not a learned model of where the real network will fail.

A.2 An explicit separation, with an equally cheap explanation

Let $n \geq 4$, $g \equiv 0$, and observe every two-element mask and $[n]$ as negative; the empty mask is known negative. There are $q = \binom{n}{2} + 1$ nonempty observations. Put $H = 2^n - \binom{n}{2} - 2$. The B_+ completion has H positives. Each singleton has $n - 1$ descending edges into negative pairs, and each coatom has one descending edge into the full mask. Thus it has $n(n - 1) + n = n^2$ descents and

$$\underline{\beta}_{S_0} = \frac{H}{2n}.$$

The other two backgrounds have zero error. Now label the unobserved singletons negative. These labels are available in S_1, because every singleton neighbors an observed pair. Keep all masks of size $3, \ldots, n - 1$ positive and the full mask negative. This explicit response is

$$B_{\text{closure}}(A) = \mathbf{1}\{2 < |A| < n\}.$$

It has $H - n$ positives and exactly n descents, all entering the full mask. Hence

$$\underline{\beta}_{S_1} \;\geq\; \frac{H - n}{2}, \qquad \frac{\underline{\beta}_{S_1}}{\underline{\beta}_{S_0}} \;\geq\; n\left(1 - \frac{n}{H}\right).$$

The observed-only and expanded warnings can therefore differ by order n. The experiment finds equality with this explicit lower bound for $n = 8, 20, 26, 30$. We do not need a theorem that the expanded optimizer is always attained by this rule: the construction already proves strict separation. Crucially, it also supplies a cheap comparator that attains the reported improvement *without* solving any cut. We present this as a completion-class example, not evidence that cut optimization is necessary.

A.3 Why predicted-positive targeting is a strong baseline

Let $p_g = \Pr_\nu[g = 1]$ and write $e_+ = \Pr_\nu[g = 1, B = 0]$, $e_- = \Pr_\nu[g = 0, B = 1]$. Uniform conditional sampling from the explanation's positive or negative regions finds a contradiction with probability e_+/p_g or $e_-/(1 - p_g)$, respectively, when the denominators are nonzero. If every minimal positive set of B has actually been found, then $g(A) = \max_{T \subseteq A} B(T)$ and $e_- = 0$: all remaining error is recovery inside the predicted-positive region. In that particular case predicted-positive sampling amplifies the per-draw detection probability by $1/p_g$ compared with unconditioned sampling. This is an elementary consequence of the response representation, not a new query algorithm. When discovery is incomplete, e_- can be nonzero and a positive-only strategy can miss it. For that reason, targeted falsification never substitutes for validation under the declared target distribution.

A.4 A LOWER WARNING CAN RISE WHILE EXACT FRAGILITY DOES NOT

For a fixed g, adding a matching observation shrinks the full alternative class, so exact β_Γ cannot increase. This does *not* imply monotonicity of a recomputed lower bound: newly fixed labels and a changed active set define a different restricted portfolio. To see this, set $n = 4$, $g \equiv 0$, and observe all six pairs and the full mask as negative. The observed-only B_+ labels four singletons and four triples positive, giving eight errors and sixteen descents. Thus its ratio is one. Adding one negative singleton removes one error and three descending edges, changing the lower bound to $4 \cdot 7/(2 \cdot 13) = 14/13$.

Exact fragility is two for both records. Any consistent response can be positive only on unobserved singletons and triples. If it has s positive singletons and t positive triples, its ratio is

$$\frac{4(s+t)}{2(3s+t)} \le 2,$$

with equality for positive triples and no positive singletons. Such a response remains allowed after fixing one singleton negative. Complete enumeration checks this example. A report must therefore say "the computed lower bound increased," not "the true fragility increased." An old attaining completion may also be contradicted by the new observation; retaining its numerical value without rechecking the response is invalid. By contrast, expanding S at a *fixed* Γ, with exact outside mass, gives nested subclasses and a nondecreasing subclass optimum as in Theorem 2.

B OVERLAPPING-REFERENCE PROBABILITY CALCULATIONS AND CONTROLS

B.1 A CLASSICAL REFERENCE-ONLY IMPORTANCE SAMPLER

For a product-mask distribution and nonempty reference terms $e_1, \dots, e_s$, write $w_j = p^{|e_j|}$, $W = \sum_j w_j$, and $\mu = \Pr_\nu[g = 1]$. Select term j with probability w_j/W, then sample a mask conditioned on containing e_j. If $c(A)$ is the number of reference terms contained in the sampled mask, set $Y = 1/c(A)$. This is the classical union/DNF importance-sampling construction (Karp et al., 1989). Its mixture probability is

$$Q(A) = \nu(A)c(A)/W, \quad \text{and hence} \quad \mathbb{E}_Q[WY] = \sum_{A:g(A)=1} \nu(A) = \mu.$$

Because $1/s \le Y \le 1$, M independent draws and Hoeffding's inequality give, with probability at least $1 - \alpha$,

$$\left|\mu - W\overline{Y}\right| \le W(1 - 1/s)\sqrt{\frac{\log(2/\alpha)}{2M}}.$$

We intersect this interval with the deterministic bounds $\max_j w_j \le \mu \le \min(1, W)$. The empty reference is handled exactly, and $s = 1$ has zero sampling width. Computing $c(A)$ costs $O(s)$ term checks; no network label is queried and the full cube is not enumerated. This is not claimed to be the sharpest available approximate-counting method.

Suppose $L \le \mu \le U$ on this single event. For background zero use

$$L_{\text{out}} = \max\left(0, L - \sum_{A \in S:g(A)=1} \nu(A)\right),$$

and for background one use

$$L_{\text{out}} = \max\left(0, 1 - U - \sum_{A \in S:g(A)=0} \nu(A)\right).$$

Every such expression is a lower bound on the actual outside disagreement mass. The same event supports every active set selected subsequently, even adaptively, for the *same* fixed reference and distribution: the inequality is deterministic conditional on $L \le \mu \le U$. There is no additional multiple-testing factor for the selected cuts or candidate masks. Changing the reference or distribution requires a new confidence allocation. In the stress test, 18 fixed reference/distribution pairs each

receive $\alpha = 0.01/18$ and $M = 16{,}384$, giving at least 99% joint coverage for the reference-counting intervals. This probability is separate from neural-label uncertainty and from behavioral-validation confidence.

For rational $p = a/d$, interval endpoints are rounded outwards to integer weights with denominator d^n. The implementation includes conservative floating rounding slack when constructing the interval; it is not a machine-checked real-arithmetic proof. All subsequent boundary and cut weights are arbitrary-precision integers. Small-instance tests check the bounds against exact weighted counts and the supplied-mass path against full response evaluation.

B.2 Stress-test design and interpretation

The three overlapping references are cyclic adjacent triples, triples sharing a common coordinate with adjacent leaf pairs, and n uniformly generated distinct triples with a fixed seed. The units are $n \in \{20, 26, 30\}$ and the product probabilities are $p \in \{1/2, 1/10\}$. Both record types include each minimal triple's positive label, its negative pair deletions, and one randomly grown maximal harmless mask. The boundary-enriched type additionally observes every mask of size $n - 2$. No discovery efficiency is inferred from supplying these records.

There are $3 \cdot 3 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 144$ warning calculations: unit count, reference type, probability, record type, mass method, and active radius. The conservative union bounds and the reference interval use identical records. We retain all results, including zero components and cases with no S_1 gain. The full portfolio is never zero or infinite on these 144 calculations. Individual constant-background union bounds are often zero even though another portfolio component is positive; the reference interval removes those zero components in the tested cases. This distinction prevents presenting a stronger lower bound as rescuing a previously vacuous *full* report.

The ordinary stress records need not identify g among all monotone functions. Their exact unrestricted fragility can therefore be infinite even when the returned subclass value is finite. These tests evaluate reference counting and supported lower bounds, not approximation ratios to a finite full optimum. Reported runtimes describe these executions and are not controlled throughput comparisons.

Under sparse product measures, very large coefficients can reflect the edge distribution rather than complicated computation. The simple one-minimal-set response $B_i(A) = \mathbf{1}\{i \in A, A \neq [n]\}$ has

$$e = p - p^n, \qquad \eta = \frac{n-1}{n} p^{n-1}, \qquad \frac{e}{\eta} = \frac{n}{n-1}(p^{2-n} - p).$$

Small p makes its only descents—near the full mask—extremely unlikely. A huge coefficient under this design is not a huge observed error or evidence that the network follows this constructed response.

C Prompt and mask-pair uncertainty in GPT-2 measurements

C.1 Two different uncertainties

Both uncertainty analyses reuse the same GPT-2 predictions; resampling does not provide additional model evaluations or prompts. There are 32 fixed prompts from four templates, 1,018 distinct masks, and 256 adjacent-pair draws at each of two product densities. For fixed prompts, pair-level binomial intervals quantify Monte Carlo uncertainty over the mask-pair distribution. They do not quantify uncertainty over prompts. The prompt analysis performs 4,000 paired prompt-bootstrap replicates, drawing eight prompts with replacement within each template and applying the *same* weights to all masks. It recomputes discrepancy, harmful-lower status, and recovery after resampling. Template-stratified intervals condition on these four templates and their fixed proportions; they do not validate a broader IOI population.

At the primary uniform-pair setting, the two recovery events add heads $(2, 2)$ and $(10, 6)$. Their smaller-to-larger discrepancy changes are respectively $4/32 \to 2/32$ and $7/32 \to 3/32$. The paired bootstrap intervals for the decreases are $[0, 5/32]$ and $[0, 8/32]$. These intervals include zero; the selected events are not independently confirmed population-level recovery effects. The intervals are descriptive and are not corrected post-selection hypothesis tests. In the published classification, these are a Previous Token head and a Backup Name Mover head, respectively; neither is one

Table 4: Conditional recovery among sampled pairs whose smaller mask is harmful. The last column is a paired template-stratified prompt-bootstrap interval for that conditional rate, in percent. Both uncertainty analyses use the same fixed prompts and zero-ablation predictions.

p, τ	Descents / pairs	Harmful lowers	Conditional %	Prompt-bootstrap %
0.2, 0.05	2/256	98	2.04	[0.00, 3.49]
0.2, 0.10	0/256	71	0.00	[0.00, 2.60]
0.2, 0.25	0/256	53	0.00	[0.00, 9.09]
0.5, 0.05	2/256	235	0.85	[0.43, 2.63]
0.5, 0.10	2/256	218	0.92	[0.46, 3.79]
0.5, 0.25	5/256	172	2.91	[0.61, 4.62]

of the two Negative Name Movers (Wang et al., 2023, Figure 20). This attribution follows that paper's classification and is not inferred from our recovery events. The added-head analysis includes zero-event heads and counts of harmful lower masks.

C.2 What an evidence-only report returns

The observed record includes observed positive-to-negative comparable pairs at each reported threshold. Thus it is inconsistent with *any* monotone response, regardless of a proposed minimal-group family. The generated evidence report records $n = 26$, the 1,017 nonempty query masks, the zero patch policy, the fixed prompt population, each threshold, and explicit contradiction witnesses. It rejects the monotone class; the inferred group family, conditional rank, and compatible-reference warning are left unset. This is the correct output for this record, not a missing numerical coefficient to fill in by ignoring contradictions. No independent behavioral validation of a discovered pretrained explanation is claimed.

D Scope of the compact neural realization

The following small threshold circuit realizes the ablation truth function. The unablated hidden values do not depend on the input, so it supplies no evidence about representation learning, meaningful IOI computation, or the prevalence of recovery in trained networks. The neural realization is explicit. Let $h_j = 1$ when unit j is retained and zero when ablated. The clean output is negative, and threshold the logit

$$z(h) = \tfrac{1}{2} - h_i - \mathrm{ReLU}\left(1 - \sum_j h_j\right). \tag{13}$$

It changes sign exactly when unit i is removed but some unit remains. Thus the obstruction is realizable by a small network, not only an arbitrary truth table. It establishes possibility within this simple architecture, not frequency in learned models. Here "rank one" describes the minimal-positive family; it does not mean a monotone rank-one DNF represents this nonmonotone response.

E An anytime report of conditional rank coverage

E.1 Proof of Proposition 1

Write $\rho_{\mathcal{K}} = \max_{e \in \mathcal{K}} |e|$, with zero for an empty family. Verified minimality means that e has an observed positive label and that each $e \setminus \{i\}$ lies in an observed harmless mask. Also assume that all recorded positive masks contain a member of $\mathcal{K}$ and that $g = B_{\mathcal{K}}$ matches every recorded label.

For $r \geq \rho_{\mathcal{K}}$, consider any monotone rank-at-most-r alternative B'. If $g(A) = 1$, some verified $e \subseteq A$ is positive under B', so $B'(A) = 1$. If $B'(A) = 1$ while $g(A) = 0$, choose a minimal positive $T \subseteq A$ of B'. Then $|T| \leq r$, $g(T) = 0$, and T cannot lie in any observed negative N. Therefore T is feasible in (4). When $r < t_*$ no such difference is possible, proving identification for the reported interval of ranks.

Conversely, a feasible minimizer T gives $B'(A) = g(A) \vee \mathbf{1}\{T \subseteq A\}$. It preserves every observed negative because no such mask contains T, and preserves all observed positives because g already explains them. It differs at T. A feasible T cannot be a strict subset of a verified e: selecting $i \in e \setminus T$ would put T in a known negative superset of $e \setminus \{i\}$. Consequently the enlarged minimal family has rank at most $\max(\rho_{\mathcal{K}}, t_*)$ and is distinct. In particular, if $t_* \leq \rho_{\mathcal{K}}$, no rank class containing g has yet been identified. The report correctly leaves the certified interval empty instead of claiming rank R for a reference of higher rank.

If no feasible T exists, every possible extra monotone term is excluded and $R = n$. If observations matching a fixed g are added, every old negative constraint remains and new ones can only shrink the feasible set; R cannot decrease. This is a deterministic statement at every completed evaluation stage. It does not require a statistical stopping rule when the labels are truthful. A new discovered edge can change the reference; the checks and interval must then be recomputed for that reference.

E.2 Integer formulation and solver limits

Use n binary variables $x_i = \mathbf{1}\{i \in T\}$ and minimize $\sum_i x_i$ subject to

$$\sum_i x_i \geq 1, \qquad \sum_{i \in e} x_i \leq |e| - 1 \quad (e \in \mathcal{K}), \qquad \sum_{i \notin N} x_i \geq 1 \quad (N \in \mathcal{N}).$$

The first constraint excludes the empty set, the second prevents containing an existing edge, and the third requires escaping every negative observation. No constraint restricts $|T|$ to b: a non-queryable unresolved candidate is precisely an intervention-ceiling limitation that should remain visible. The instance has n variables and $1 + |\mathcal{K}| + |\mathcal{N}|$ constraints. It avoids a full response table but can still be computationally hard. We use HiGHS through SciPy, a zero requested mixed-integer gap, and an explicit time limit. Only a successful optimum or proved infeasibility produces R; other statuses leave the optimization unresolved. Returned feasible masks are checked independently. Solver status is not a separately machine-checked formal optimization proof.

E.3 Report fields and appropriate decisions

Given an intervention record, the report specifies $\mathcal{K}, \mathcal{N}$, the observed labels, found rank, resolved-rank interval, b, τ, replacement policy, and a minimum unresolved mask when one exists. It also returns a boundary-complete warning, its full response representation, queryable candidate falsifications, and a separate behavioral-validation field. The latter states that validation has not been performed unless a genuine independent validation has been supplied. A finite warning is not an acceptance certificate; an infinite warning can simply reflect unresolved higher-order monotone alternatives.

The default practical sequence is to retain the discovery record, check its logical consistency, state the rank classes it resolves, and use any budget for additional queries or an independent behavioral check. A discovered contradiction ends support for that *fixed* explanation. A rank claim always reads "identified within the monotone rank-at-most-r class," not "the true response has rank r." If a practitioner needs the latter, architectural evidence or a richer verification method is necessary. Our report does not infer a true rank upper bound from the absence of observations.

For finitely sampled query labels, all deterministic report statements hold on any simultaneous event on which every released label is truthful, including after adaptive stopping. Independently, if a sequence of frozen explanations is directly validated at times $j = 1, 2, \ldots$, use fresh mask draws and one-sided binomial intervals with error allowances $\alpha_j = 6\alpha/(\pi^2 j^2)$. Conditional validity at each time and a union bound give total false acceptance probability at most α. This standard allocation is a supporting option, not a new confidence-sequence result. Reusing a single holdout to adapt the explanations is not covered.

F BOUNDARY-COMPLETE SPARSE WARNINGS

F.1 FULL-DOMAIN DECOMPOSITION AND PROOF OF THEOREM 2

Let $\mathcal{E} = \{(A, A \cup \{i\}) : i \notin A\}$ denote the directed cube edges. For the response (8), its exact error is

$$e(B_x, g) = E_{\text{out}} + \sum_{A \in S} \nu(A) \mathbf{1}\{x_A \neq g(A)\}.$$

Its exact local-violation probability is

$$\begin{aligned}\eta(B_x) = &\sum_{(A,A') \in \mathcal{E}:\ A, A' \in S} \pi(A, A') x_A (1 - x_{A'}) \\ &+ \sum_{(A,A') \in \mathcal{E}:\ A \in S, A' \notin S} \pi(A, A') x_A (1 - h(A')) \\ &+ \sum_{(A,A') \in \mathcal{E}:\ A \notin S, A' \in S} \pi(A, A') h(A) (1 - x_{A'}).\end{aligned}$$

Edges with both endpoints outside S have zero contribution because their labels come from monotone h. Thus every potentially nonzero edge is either internal or crosses the boundary. There are at most $n|S|$ such edges to inspect, with duplicate internal edges counted only once.

For fixed λ, maximizing $e - \lambda\eta$ is a constant plus unary terms and the directed submodular pair costs $-\lambda\pi(A, A')x_A(1 - x_{A'})$. Unary disagreement is encoded as in the full graph. A boundary edge to a fixed outside zero becomes an active-to-sink arc; a boundary edge from an outside one becomes a source-to-active arc. Observed labels are fixed by capacities exceeding the sum of all finite capacities. Consequently every cut gives a legitimate full response matching the observations, and standard binary-energy minimization applies (Kolmogorov & Zabih, 2004).

Fractional updates (Dinkelbach, 1967) start at ratio zero. A maximizing response with positive error and zero descent mass proves infinite subclass fragility. Otherwise a positive residual objective yields a strictly larger ratio; a nonpositive maximum certifies the current optimum. There are only finitely many active labelings. Our implementation uses exact integer arithmetic for rational product probabilities, returns the last attaining labeling, and raises an unresolved-computation error at a declared iteration cap instead of claiming an unproved optimum. The number of fractional updates is not claimed strongly polynomial. Each cut graph, however, is polynomial in the explicitly supplied active set.

The optimized class is a subset of all full responses consistent with Γ, so its ratio cannot exceed β_Γ. If $S \subseteq S'$, every old extension is represented in the larger class by fixing newly active nodes to their old background labels; the exact subclass optimum cannot decrease. If only $L_{\text{out}} \leq E_{\text{out}}$ is available, each surrogate numerator is at most the true error of the same full response, while its denominator remains exact. Optimizing the surrogate therefore remains a valid lower bound, although it need not equal the exact subclass optimum. A positive surrogate error and zero denominator still proves infinite full fragility.

F.2 RATIONAL PRODUCT MEASURES WITHOUT HIDDEN ENUMERATION

Let the ablation probability be $p = a/d \in (0, 1)$ in lowest terms and put $c = d - a$. A mask A has integer weight $a^{|A|}c^{n-|A|}$ with denominator d^n. An upward edge with lower endpoint A has integer weight $a^{|A|}c^{n-1-|A|}$ with denominator nd^{n-1}. If H and V are the weighted error and descent sums, respectively, the ratio is $(n/d)H/V$. Integer capacities can be large; we use an arbitrary-precision cut implementation for this sparse calculation, not fixed-width capacities that could overflow at large n.

For $h = g$, outside error is exactly zero. For a constant background, its total error is the probability that the reference takes the opposite value. For disjoint minimal groups e, independence gives

$$\Pr[g(A) = 0] = \prod_{e \in \mathcal{K}} (1 - p^{|e|}).$$

This includes unused coordinates without further computation. For large overlapping groups, default bounds are

$$\max_{e\in\mathcal{K}} p^{|e|} \;\le\; \Pr[g=1] \;\le\; \min\left(1, \sum_{e\in\mathcal{K}} p^{|e|}\right).$$

A lower bound on total error minus the exactly counted in-S background error, clipped below at zero, is a lower bound on E_{out}. The bounds can be loose; a zero result is not evidence of low fragility. Exact DNF counting is permitted only when explicitly requested, or automatically for small systems/disjoint groups, and is not advertised as polynomial for arbitrary overlaps. The overlapping stress tests in Appendix B compare union bounds with a simultaneous reference-mass interval.

For S equal to observed masks and their radius-one Hamming neighbors, $|S| \le (q+1)(n+1)$. The construction does not require S to be a sublattice. Closing under unions and intersections is unnecessary and potentially exponential.

F.3 A COUNTEREXAMPLE TO DROPPING BOUNDARY EDGES

Take $n = 4$, $g \equiv 0$, and only the full-mask negative query. The full coefficient is seven by (20). Set S to the empty mask, all four singletons, all four triples, and the full mask. This is precisely the radius-one neighborhood of the anchors. A partial response positive on the singletons and negative on the remaining nodes has positive numerator but no descent on a retained adjacent edge. Its induced-subgraph ratio is infinite under either raw positive weights or renormalized retained weights. Since the full coefficient is finite, this is not a valid lower bound. Extending the partial labels is possible, but the extension must introduce descents on ignored edges. The existence of an extension does not preserve its partial ratio. An exact optimization on a small subset is therefore useful only after its full extension and boundary costs have been specified.

F.4 PRODUCING AN ACTUAL QUERY WITHOUT ENUMERATING MASKS

The finite-candidate procedure scans active nodes, their Hamming neighbors, and the reported minimal groups. It returns unobserved masks where the explicit completion disagrees with g. For the supported backgrounds, a positive conservative numerator guarantees that this search finds a mask.

For $h = g$, every disagreement is already active. For $h = 0$, an outside disagreement has $g = 1$ and thus contains a minimal group. Either that group is already outside S, or an upward path from an active group to the outside positive mask crosses the boundary at another positive mask. For $h = 1$, an outside disagreement lies in the harmless down-set of g; an upward path from the known empty mask to it stays harmless and crosses the boundary. Thus a witness can be exhibited with $O(n|S| + |\mathcal{K}|)$ candidate evaluations, not a cube enumeration. Each candidate can be checked against b; an out-of-budget candidate is not silently treated as queryable. This procedure is for falsification, not an independent sample from ν.

The tiny-model study evaluates this finite-candidate procedure separately from a full-support comparator that samples uniformly from the entire disagreement support, enumerable at eight heads. Their candidate domains differ, so full-support detection rates do not establish the performance of the finite-candidate procedure.

G RESTRICTED ALTERNATIVES AND SHARP COEFFICIENTS

G.1 PROOF OF THEOREM 4

For $g \equiv 0$, the error count is the number H of positive masks. Let V be the number of descending upward edges; the uniform ratio is $(n/2)H/V$. Any nonzero response with at most one inclusion-minimal positive set has a unique such set $T \neq \varnothing$. Every positive mask contains T: descending through its finite set of positive subsets must reach that same minimal element. Write $d = n - |T|$. Since the full mask is fixed negative, $1 \le d \le n-1$. All positives lie in the d-dimensional face of supersets of T.

Send one unit of flow from every positive mask to the full mask along a uniformly random upward path in that face. Each path crosses at least one descent, because it begins positive and ends negative.

Upper-bound loads by sending unit flow from every vertex of the face except its top, whether positive or not. By symmetry an edge at face level j has load

$$L_j = \frac{\sum_{\ell=0}^{j} \binom{d}{\ell}}{(d-j)\binom{d}{j}}, \qquad 0 \le j \le d-1.$$

The denominator is $d\binom{d-1}{j} \ge d$, and the numerator is at most $2^d - 1$. Hence each descending edge carries at most $(2^d - 1)/d$ units, so

$$H \le V \frac{2^d - 1}{d}, \qquad \frac{e}{\eta} \le \frac{n(2^d - 1)}{2d}.$$

The sequence $(2^d - 1)/d$ is strictly increasing for integers $d \ge 1$, because the numerator of its next difference is $(d-1)2^d + 1 > 0$. Taking $d = n - 1$ proves the upper bound in (10).

For equality, set $T = \{i\}$ and label precisely the sets A satisfying $T \subseteq A \subsetneq [n]$ positive; label everything else negative. This is exactly $B_i(A) = \mathbf{1}\{i \in A, A \ne [n]\}$. Its positive count is $2^{n-1} - 1$ and its only descents are the $n - 1$ edges from a coatom containing i into the full mask. Its sole minimal positive is a singleton. On every increasing maximal chain it changes $0 \to 1$ when i is added, unless i is last, and changes back only at the full mask. Thus it has at most one descent on each chain. The construction remains in the additionally restricted class, proving equality there too.

For the neural realization (13), all hidden intervention units equal one before ablation. If i is retained, the logit is at most $-1/2$. If i is removed but some unit remains, the ReLU is zero and the logit is $1/2$. If every unit is removed, the ReLU is one and the logit is $-1/2$. The clean prediction is negative. Zero-ablation disagreement therefore equals B_i exactly, with a nonzero margin in every case. The construction is a small thresholded ReLU network, not evidence that trained transformers typically implement this rule.

G.2 A BOUND ON THE TOTAL NUMBER OF DESCENTS

Within the same one-minimal-set class, suppose the total number of descending edges is at most k. The face above T has d edge-disjoint upward paths from its bottom to its top. One construction adds the d free coordinates in their d cyclic orders: at an intermediate cardinality each cyclic interval is different, so the paths share no directed edge. Every path starts at the positive T and ends at the negative full mask, hence contains a descent. Therefore $V \ge d$ and $d \le k$.

The preceding flow bound is increasing in d, so the ratio is at most $n(2^{d_*} - 1)/(2d_*)$ with $d_* = \min(k, n-1)$. Choose a fixed nonempty T of size $n - d_*$ and label precisely the sets A satisfying $T \subseteq A \subsetneq [n]$ positive. It has $2^{d_*} - 1$ positives and d_* descents, attaining the bound. This proves (11). With zero allowed descents only the all-negative response is consistent, so the coefficient is zero. A total-descent restriction is a global combinatorial restriction, not something inferred from a handful of observed adjacent pairs. We do not assert the same sharp formula for arbitrary responses with many minimal positive sets.

G.3 MULTIPLE LARGE-MASK QUERIES STILL PERMIT A SIMPLE OBSTRUCTION

The one-query example extends to some larger records without allowing complex minimal-positive families. Let all q observed nonempty masks be negative, include the full mask, and have size at least $n - c \ge 2$. For fixed i, let $\mathcal{Z}_i$ contain those observed masks containing i and define

$$B_i(A) = \mathbf{1}\{i \in A\} \prod_{N \in \mathcal{Z}_i} \mathbf{1}\{N \not\subseteq A\}.$$

This matches all observations. Its sole minimal positive is $\{i\}$, and along any increasing chain it can recover at most once: once a blocking N is completed, it stays completed. Let $S_i = \sum_{N \in \mathcal{Z}_i} 2^{-|N|}$. Under uniform masks, a union bound gives $e(B_i, 0) \ge 1/2 - S_i$. A descent must add a coordinate $j \ne i$ that completes a blocking mask, so

$$\eta(B_i) \le \frac{2}{n} \sum_{N \in \mathcal{Z}_i} (|N| - 1) 2^{-|N|}.$$

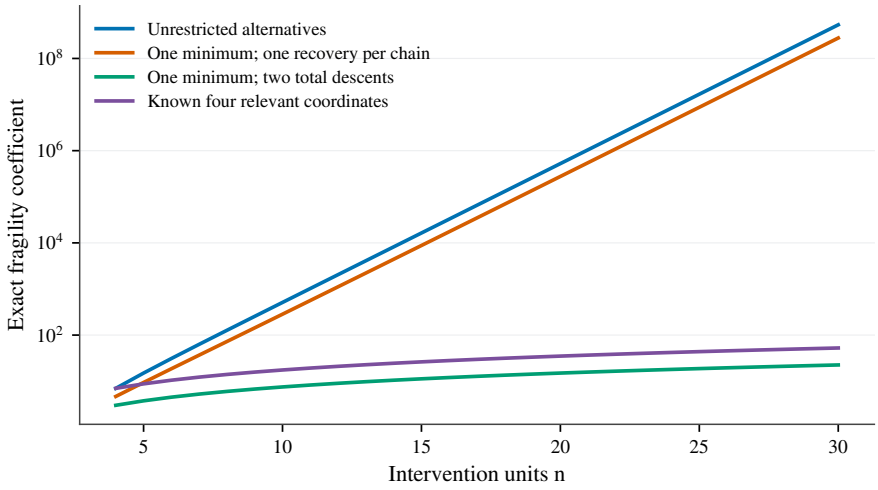

Figure 1: Exact formulas, not fitted experimental curves, for an all-negative reference with only a full-mask negative observation. One recovery per chain and one minimal positive set still permit exponential fragility. In contrast, two total descents in the one-minimal-set class, or four supplied relevant coordinates, give linear dependence on the ambient dimension. These are different declared alternative classes, not facts inferred from the query.

Whenever the numerator bound is positive, this explicit response proves

$$\beta_\Gamma^{\text{one, chain}\le 1} \ge \frac{n(1/2 - S_i)}{2\sum_{N\in\mathcal{Z}_i}(|N|-1)2^{-|N|}} \ge \frac{n(2^{n-c}-2q)}{4q(n-1)}. \tag{14}$$

This preserves an exponential-in-$n-c$, inverse-q obstruction for this class of negative, large-mask records. It is not a claim for every arbitrary q-query transcript. A network realization replaces the single ReLU in (13) with $\sum_{N\in\mathcal{Z}_i}\mathrm{ReLU}(1-\sum_{j\in N}h_j)$, using $O(nq)$ connections. The construction therefore need not encode an exponential truth table in its architecture.

G.4 Known effective coordinates and nonuniform measures

Assume that g and all alternatives depend only on a supplied coordinate set J, $|J|=j$. A matching record projects consistently onto 2^J. Every projected Boolean completion has exactly one lift in the restricted class. Under uniform masks its error is unchanged by the irrelevant coordinates. An adjacent edge changes a relevant coordinate with probability j/n; conditional on that event its projection is a uniform j-cube edge. Thus $\eta_n = (j/n)\eta_j$ and taking the supremum gives (12). This is an exact reduction for a known coordinate restriction, not a junta-identification algorithm or a claim about learning J from interventions.

For the explicit singleton witness B_i under product masks with ablation probability p, the exact quantities are

$$e_p(B_i,0) = p - p^n, \qquad \eta_p(B_i) = \frac{n-1}{n}p^{n-1}, \qquad \frac{e_p}{\eta_p} = \frac{n}{n-1}(p^{2-n}-p).$$

Lower-density masks do not automatically reduce pessimism: they can make the rare top-boundary descents even rarer. These are distribution-specific ratios, not upper bounds on learned-model error. The known-coordinate example, in contrast, gives an explicit setting where stronger structural information substantially reduces the *optimal* coefficient.

H Definitions, edge cases, and proof of exact identification

We work on $2^{[n]}$ with $n \ge 1$, $1 \le r, b \le n$, and a known harmless empty mask. The empty failure family is allowed; an empty failure *edge* is not. A transcript records distinct nonempty queried

masks and their truthful labels. Counting repeated calls instead can only increase a method's cost. A certificate identifies a target among all monotone rank-at-most-r responses, with no supplied bound on their number of minimal edges. The target-dependent quantities s, ρ, α, C are not inputs to discovery.

Write $\mathcal{V}_r = \{T : 1 \le |T| \le r\}$ and $\mathcal{I}_r(\mathcal{H}) = \{T \in \mathcal{V}_r : B_{\mathcal{H}}(T) = 0\}$. Put $s = |\mathcal{H}|$, $\rho = \mathrm{rank}(\mathcal{H})$, and let $\alpha(\mathcal{H})$ be the largest size of a harmless mask. The negative-cover cost $C_{b,r}(\mathcal{H})$ is the minimum number of harmless allowed masks covering every member of $\mathcal{I}_r(\mathcal{H})$ by inclusion. It is zero for an empty candidate set and infinite if no cover exists. The smallest sufficient intervention-size ceiling is called the *identifiability radius*; it refers to mask cardinality, not a metric perturbation radius.

Proposition 5 (Identifiability and certificate size)**.** *The full response table on masks of size at most b identifies a target $\mathcal{H}$ among rank-at-most-r monotone responses if and only if*

$$b \ge b_* := \max\{\rho, \min(r, \alpha(\mathcal{H}))\}. \tag{15}$$

When this holds, the smallest truthful identifying transcript has exactly

$$\mathrm{Cert}_{b,r}(\mathcal{H}) = C_{b,r}(\mathcal{H}) + s \tag{16}$$

nonempty queries. Otherwise no finite allowed identifying transcript exists.

H.1 Proof of Proposition 5: radius

If a true minimal edge e has $|e| > b$, the antichain $\mathcal{H} \setminus \{e\}$ gives the same responses on all allowed masks. It differs at e, so identification is impossible. If instead $r > b$ and $\alpha(\mathcal{H}) > b$, choose a harmless set T of size $b + 1$. The function

$$B'(A) = B_{\mathcal{H}}(A) \vee \mathbf{1}\{T \subseteq A\}$$

also has a minimal antichain of rank at most r, agrees on every allowed mask, and differs at T. Removing subsumed edges when taking the minimal antichain does not increase rank. These constructions prove necessity of both terms in (15).

Conversely, suppose $b \ge \rho$ and $b \ge \min(r, \alpha)$. Every true edge is queryable. Every harmless candidate $T \in \mathcal{V}_r$ has $|T| \le \min(r, \alpha) \le b$ and is also queryable. A consistent alternative must keep the true edges positive and all these candidates negative. If the alternative differed on a positive candidate of the target, it would contradict a true-edge witness. If it differed on a harmless candidate, it would contradict that candidate's queried label. A rank-at-most-r monotone function is determined by its labels on $\mathcal{V}_r$: any positive larger set contains a positive candidate of size at most r. Thus the full function is identified.

For the empty family, $\rho = 0$ and $\alpha = n$, so the radius is r. For a function positive on every nonempty mask, $\alpha = 0$, $\rho = 1$, and the radius is one. These cases are included in the exhaustive checks.

H.2 Proof of Proposition 5: certificate size

Suppose a harmless candidate $T \in \mathcal{I}_r$ lies in no queried negative mask. Adding its positive term, as above, cannot change any observed negative label, because such a change would require T to be contained in that negative query. It cannot change a positive label either. Therefore each harmless candidate must be covered by an observed negative superset, requiring at least C negative queries.

For every true minimal edge e, at least one positive query must contain e but no other true edge. Otherwise deleting e would preserve the transcript. One query can isolate at most one true edge in this sense. Hence at least s positive queries are required, independently of the negative-cover requirement.

For sufficiency, query the s true edges and any optimal negative cover of size C. An alternative consistent response must keep every true edge positive. Every harmless candidate must remain negative because it is contained in an observed negative query. This fixes all labels in $\mathcal{V}_r$, and consequently the entire response. The transcript has $C + s$ queries. Such a certificate is target-dependent, as a teaching set is (Goldman & Kearns, 1995); the theorem does not assume that an uninformed procedure can choose it directly.

The independent-cover lower bound is already present in the hidden-hypergraph literature (Angluin & Chen, 2006). Our statement specifies the intervention ceiling and declared rank and combines negative

coverage with the positive-witness requirement. A supplied tight edge-count bound invalidates some adding-edge alternatives and is intentionally excluded.

H.3 Why the residual checker is sound, conditionally

The checker requires each reported edge $e \in \mathcal{K}$ to have an observed positive witness at e (or a validated equivalent inferred positive), and for each $i \in e$ it requires $e \setminus \{i\}$ to be contained in an observed negative mask. Monotonicity then proves that e is positive and every one-element deletion is negative. Every proper subset is contained in a one-element deletion, proving minimality. Every observed positive query must contain a found edge. Positive-subset/negative-superset contradictions are rejected.

If $\mathcal{U} = \varnothing$, every candidate either contains a verified edge and is positive, or lies inside an observed negative and is negative. This determines the response in the declared class. If $T \in \mathcal{U}$, the two monotone functions $B_{\mathcal{K}}$ and $B_{\mathcal{K}} \vee \mathbf{1}\{T \subseteq A\}$ both match the transcript and differ at T. In particular, a residual T cannot be a strict subset of a verified edge: a deletion witness for that edge would have excluded it. Minimizing the enlarged antichain preserves rank at most r. Hence the residual also gives a constructive obstruction to completeness at a finished discovery stage.

These logical checks do not inspect unqueried labels. A nonmonotone oracle can pass them. An internal algorithm status is weaker still: our transformer results discard internal completion flags that fail the separate checker.

H.4 The response encoded by all minimal failures

Let B be any Boolean response with harmless empty mask, without monotonicity. Every positive mask contains an inclusion-minimal positive subset, because the subset lattice is finite. Therefore its minimal-positive family represents

$$G_B(A) = \max_{T \subseteq A} B(T).$$

This is monotone and dominates B. Any other monotone response dominating B must be positive on every superset of each positive mask, so it must dominate G_B. Hence G_B is the least monotone upper envelope. It is not necessarily the monotone approximation with smallest average error, since such an approximation may sacrifice true positive masks.

Increasing-order enumeration correctly returns the minimal-positive family even for a nonmonotone response. Any queried positive has no previously found positive subset; if it had a positive proper subset, an inclusion-minimal one would already have been encountered. Conversely, a true minimal positive is not skipped by a found smaller edge. The final representation is nevertheless G_B, not necessarily B. With full-rank/full-radius access and the shared cached implementation, its distinct nonempty query count is $|\{A : G_B(A) = 0\}| - 1 + |\mathcal{H}|$: every nonempty negative mask of the closure is queried, each minimal positive is queried, and positive supersets are skipped. A test checks this identity and family correctness for all 128 three-unit Boolean responses with harmless empty mask. This explains why exact family recovery by the increasing-order baseline does not resolve the transformer representation errors.

I Certificate-relative bound for an exponential-time discovery rule

Proposition 6 (Global query bound)**.** *Assume* (15)*, exact labels, and exact coverage maximization. Let* $d = \min\{|\mathcal{I}_r|, \sum_{j=1}^{\min(r,b)} \binom{b}{j}\}$ *and* $H_d = \sum_{j=1}^{d} 1/j$*, with* $H_0 = 0$*. The procedure recovers and certifies* $\mathcal{H}$ *using*

$$Q \le H_d C_{b,r}(\mathcal{H}) + (b+1)s. \tag{17}$$

I.1 Algorithm and correctness

Initialize the found family and negative ledger to empty, with $\mathcal{U} = \mathcal{V}_r$. Among masks with $|A| \le b$ containing no found edge, choose one maximizing $|\{T \in \mathcal{U} : T \subseteq A\}|$. A negative answer adds A to

the ledger. A positive answer is shrunk by considering each component once: remove it when the remaining mask is still positive, otherwise keep it and record the negative deletion. Record the final positive edge and all negative evidence, and update (3).

If a deletion was negative when considered, its later smaller version is still negative under monotonicity. The final positive mask is therefore minimal. It is new because the original pool avoided all previously found edges. Any residual positive candidate contains an undiscovered true edge, which is queryable under (15) and supplies a feasible positive-score pool. Any residual negative candidate is covered by a true negative allowed mask, which remains feasible and has positive score. Thus a nonempty residual cannot stall under the theorem's assumptions. Each query or completed extraction removes candidates, so the finite process terminates and the residual checker certifies the true family.

When the radius condition fails, the implementation preserves unqueryable residual candidates instead of declaring success. A supplied query limit gives a budget-limited result. Uncertain labels or detected assumption contradictions likewise stop without a checked certificate.

I.2 Proof of Proposition 6

Fix an optimal true-negative cover $\mathcal{F}$ of size C. Assign each member of $\mathcal{I}_r$ to one covering mask in $\mathcal{F}$, producing disjoint blocks P_A with $|P_A| \leq d$. All masks in $\mathcal{F}$ remain feasible throughout discovery: a true-negative mask contains no true edge and hence no found edge.

Consider an outer iteration whose answer is negative. Let k be its number of newly covered residual candidates. All k are true negatives. If some block P_A has q_A unresolved elements before this iteration, exact coverage maximization implies $k \geq q_A$, since the fixed mask A is still feasible. Charge $1/k$ to each newly covered candidate. The total charge of this outer query is one. If it covers t elements of a block with q_A remaining, that block is charged at most

$$\frac{t}{q_A} \leq \sum_{j=q_A-t+1}^{q_A} \frac{1}{j}.$$

Negative queries during extraction may remove more block elements for free in this accounting. Discovering a true edge cannot remove a true-negative candidate. Thus the total charge to block P_A, across *all* discovery phases, is at most $H_{|P_A|} \leq H_d$. The number of outer negative queries is at most CH_d.

There are at most s outer positive queries. Each needs at most b additional deletion queries. This adds at most $(b+1)s$ calls, proving (17). Caching can only reduce distinct queries. When $\mathcal{I}_r$ is empty, the cover term is zero and the extraction argument still applies. The bound is a cross-phase use of the classical harmonic set-cover argument (Chvátal, 1979), not a new approximation principle.

If the chosen score is at least $1/a$ times the maximum, $a \geq 1$, the same proof gives $aH_dC + (b+1)s$. An arbitrary heuristic does not receive a proved approximation factor merely because it performs well empirically. Since an optimal transcript has $C + s$ queries, the exact-maximization method is within $\max(H_d, b+1)$ of that transcript benchmark whenever $C + s > 0$.

J Rank uncertainty, covering arrays, and locality

For m disjoint pairs, a harmless mask selects at most one endpoint from each pair. With $b \geq m$, any negative query can be extended to a maximal negative query by choosing one endpoint of every omitted pair. Such an extension never hurts coverage. A maximal negative mask is a binary string of length m; a harmless candidate specifies values in some coordinates. Covering all candidates of order at most r is therefore equivalent to including every binary assignment on each $t = \min(r, m)$-subset of coordinates. Coverage at strength t also covers smaller assignments by extension. This proves the exact correspondence to a binary covering array (Colbourn, 2004).

When $t = m$, every one of the 2^m strings must occur, proving $C = 2^m$. For strength two, label the m coordinates by distinct $\lceil \log_2 m \rceil$-bit words. Use the all-zero row, the all-one row, every bit-position row, and its complement. Constant rows provide equal assignments to any coordinate pair. Distinct coordinate labels differ at some bit, and that row and its complement provide both

unequal assignments. This constructs at most $2 + 2\lceil \log_2 m \rceil$ rows. For any fixed t coordinates their 2^t assignments require at least 2^t rows, so $C \geq 2^t$.

When $b < m$ and $t \leq b$, a negative query covers at most $\binom{b}{t}$ of the $2^t\binom{m}{t}$ oriented t-subsets. Thus

$$C \geq \left\lceil \frac{2^t\binom{m}{t}}{\binom{b}{t}} \right\rceil .$$

For $b < t$, at least one required candidate is larger than every query and no cover exists. Positive edges also have to be queryable, so the full identifiability criterion still includes $b \geq \rho = 2$.

For comparison, a single hidden failure T of size $k > b$ is indistinguishable from the empty family. At $b \geq k$, one size-b query contains $\binom{b}{k}$ possible hidden k-sets out of $\binom{n}{k}$. An exact all-negative transcript excluding every such set requires at least $\lceil \binom{n}{k}/\binom{b}{k} \rceil$ queries, even adaptively: the all-negative branch leaves any uncovered target consistent. Independent uniform size-b pools detect a fixed target with probability $p = \binom{b}{k}/\binom{n}{k}$, hence miss it after Q pools with probability $(1-p)^Q$. This is an elementary locality check, not a novelty claim.

K Full-lattice benchmark for transcript fragility

Proposition 7 (Sharp transcript fragility)**.** *For full-support ν, π, use the coefficient in* (7)*. Then $e_\nu(B, g) \leq \beta_\Gamma \eta_\pi(B)$ is sharp whenever β_Γ is finite. It is finite exactly when Γ identifies g among* all *monotone Boolean responses. Additional truthful labels cannot increase it. For $\lambda \geq 0$, a directed minimum cut on the mask lattice computes*

$$1 - \mathrm{mincut}_\Gamma(\lambda) = \max_{B|_\Gamma = g|_\Gamma} \{e_\nu(B, g) - \lambda \eta_\pi(B)\}. \tag{18}$$

Hence β_Γ is the least λ for which the cut value is one, or is infinite if no such λ exists.

K.1 Finite and infinite coefficients

Fix a monotone g matching a truthful transcript. Both ν on masks and π on upward adjacent edges have full support. There are finitely many consistent Boolean responses. For each with positive violation rate, the error-to-violation ratio is finite. Zero violation rate is equivalent to monotonicity, because every comparable pair is joined by upward adjacent edges. Full support of ν makes the error positive for every response distinct from g. It follows that β_Γ is finite exactly when g is the only consistent monotone response. When finite it is the maximum of finitely many ratios, including zero for g; an attaining response proves sharpness. If all consistent responses equal g, the coefficient is zero.

Adding labels matching g only reduces the feasible response set, so it cannot increase the supremum. With a non-full-support ν, the appropriate finiteness condition is instead that all consistent monotone responses agree with g almost surely under ν. The exact-lattice checks use uniform full-support distributions; the sparse warning and pretrained measurement additionally use product weights.

K.2 Minimum-cut construction

Create one vertex for each mask, a source s, and a sink t. The source side of a cut encodes $B(A) = 1$. Add the following capacities for a given $\lambda \geq 0$:

- If $g(A) = 0$, add $s \to A$ of capacity $\nu(A)$; if $g(A) = 1$, add $A \to t$ of capacity $\nu(A)$.
- For every upward adjacent pair (A, A'), add $A \to A'$ of capacity $\lambda\pi(A, A')$.
- Force observed positive vertices to the source side and negative vertices to the sink side with capacities larger than one. The empty mask is forced negative.

The cut represented by g costs one, so no optimal cut can violate a forced label. Unary arcs contribute precisely the mass of masks on which $B = g$, or $1 - e_\nu(B, g)$. An upward arc crosses the cut exactly for a descent $B(A) = 1, B(A') = 0$, contributing $\lambda\eta_\pi(B)$. Thus the optimal cost is

$$\min_{B|_\Gamma = g|_\Gamma} [1 - e_\nu(B, g) + \lambda \eta_\pi(B)],$$

which proves (18). The maximum on its right is nonnegative because $B = g$ is feasible. It is zero exactly when every consistent response satisfies $e \leq \lambda\eta$. A distinct consistent monotone response keeps it positive for every λ; otherwise its zero threshold is the finite coefficient. This is a specialized binary-energy reduction using classical graph-cut machinery (Kolmogorov & Zabih, 2004).

K.3 Exact uniform-distribution implementation

Let $N = 2^n$ and $L = n2^{n-1}$ be the numbers of masks and upward edges. For a response B, write $h(B)$ for its number of disagreements with g, and $d(B)$ for its number of descents. Then

$$\beta_\Gamma = \frac{L}{N} \max_B \frac{h(B)}{d(B)} = \frac{n}{2} \max_B \frac{h(B)}{d(B)}.$$

We optimize the count ratio by fractional updates (Dinkelbach, 1967). Given a rational ratio p/v, solve a minimum cut with integer unary capacities v, upward capacities p, and anchor capacities $vN+1$. The recovered response maximizes $vh - pd$ exactly, and the cut check verifies $vh - pd = vN - \text{flowvalue}$. Start at zero. A positive optimum with $d = 0$ proves infinite fragility. A zero optimum certifies the current ratio. Otherwise update it to h/d, which strictly increases it. Since the feasible set is finite, the sequence terminates after finitely many distinct ratios. The previously returned maximizing response attains the final ratio. We explicitly raise an error on a computation cap rather than report an unproved optimum.

The implementation uses integer capacities in SciPy's directed maximum flow; no floating tolerance determines optimality. The reported floating coefficient is converted from an exact rational ratio. Its graph has $N+2$ nodes and $O(nN)$ arcs. Although each cut is polynomial in this graph size, the graph is exponential in intervention units. This exact benchmark restricts $n \leq 12$. The boundary-complete partial solver in Appendix F is separate and does not enumerate that lattice.

K.4 Additional labels from an attaining response

If a finite positive coefficient is attained by B^*, any mask with $B^*(A) \neq g(A)$ must be unqueried. Query such a mask. A label different from g rejects the reference explanation. A matching label excludes B^* from the feasible response set, while preserving the true monotone reference when it is correct. This cannot increase fragility, although another response may attain the same ratio. For infinite fragility the solver instead returns a distinct monotone witness, and the same rule applies.

L Sharp sparse-transcript obstruction

Theorem 8 (Query–fragility lower bound)**.** *Take uniform masks and uniform upward adjacent pairs. For any monotone g and a truthful transcript with $q \geq 1$ distinct nonempty queried masks,*

$$\boxed{\beta_\Gamma \geq \frac{2^n - q - 1}{2q}}. \tag{19}$$

Thus a uniform guarantee $\beta_\Gamma \leq K$ requires $q \geq (2^n - 1)/(2K + 1)$.

Proof. Let P be the observed positive masks and Z the observed negative masks, including the empty mask. Consider two consistent responses: B_- is positive exactly on P; B_+ is negative exactly on Z. Every unobserved mask is mispredicted by exactly one of these responses relative to g. Their disagreement counts therefore sum to $2^n - q - 1$. A downward edge of B_- must leave a positive anchor; one of B_+ must enter a nonempty negative anchor. Their downward-edge counts sum to at most nq. Either a nonzero error has zero violations, giving infinite fragility, or one response has disagreement-to-violation count ratio at least $(2^n - q - 1)/(nq)$. Uniform probabilities multiply this ratio by $n/2$, proving the claim. □

Theorem 8 is an unrestricted, distribution-specific obstruction, not a lower bound for direct behavioral validation. Its proof uses only two explicit completions and therefore applies to any monotone g, irrespective of rank, edge count, or overlap. If $q = 2^n - 1$, the lower bound is zero, as every label is known. If no nonempty labels are observed, a distinct consistent monotone response exists and

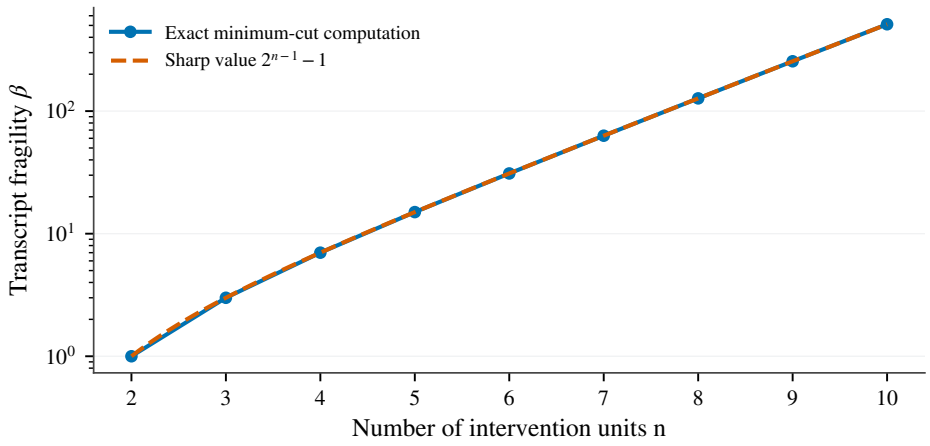

Figure 2: One harmless full-mask query exactly identifies the all-harmless monotone response, yet its sharp robustness coefficient grows exponentially. Minimum-cut values match the proved expression for every $n = 2, \ldots, 10$. This is an exact construction, not a Monte Carlo estimate.

fragility is infinite. The theorem does not assert that every particular low-query certificate attains the lower bound; it supplies a universal obstruction.

L.1 Proof of the sharp one-query value

For $g \equiv 0$ and a single harmless full-mask query, with $n \geq 2$, the exact unrestricted value is

$$\beta_\Gamma = 2^{n-1} - 1. \tag{20}$$

Let $g \equiv 0$ and force the empty and full masks negative. The response positive on all $N - 2$ nonempty proper masks has $h = N - 2$ disagreements and exactly n downward edges, all entering the full mask. Its coefficient is $(n/2)(N-2)/n = (N-2)/2 = 2^{n-1} - 1$.

For the upper bound, consider any consistent response and each of its positive masks. Send one unit of flow from that mask to the full mask along a uniformly random upward path. Every such path starts positive and ends negative, so it crosses at least one downward edge. Therefore the total positive-source count h is at most the total path flow through downward edges.

Bound this flow by sending unit flow from *every* nonempty proper mask, not just the positive ones. Symmetry gives the same load on each edge whose lower endpoint has size k. Sources of sizes $1, \ldots, k$ contribute exactly one unit across level k each, while other sources contribute none. Thus the load is

$$L_k = \frac{\sum_{j=1}^{k} \binom{n}{j}}{\binom{n}{k}(n-k)}, \qquad 1 \leq k \leq n-1.$$

Since $\binom{n}{k}(n-k) = n\binom{n-1}{k} \geq n$ and the numerator is at most $N - 2$, every edge has load at most $(N-2)/n$. Positive-source flow is no larger. With d downward edges this implies $h \leq d(N-2)/n$, hence $e/\eta = (n/2)(h/d) \leq (N-2)/2$. Equality is attained by the construction, completing the proof. Upward routing on the cube is a classical proof tool (Briët et al., 2012); the statement here concerns one fixed explanation and its transcript anchors.

L.2 Why small distance to monotonicity is not enough

The sharp witness is close to a *different* monotone function: the response positive on every nonempty mask differs only at the full mask. Its distance to that function is $1/N$. But that nearby function contradicts the observed full-mask label. Distance to the monotone class therefore cannot replace transcript-conditioned distance to g. This explains why the result does not contradict classical or modern monotonicity-testing bounds (Goldreich et al., 2000; Khot et al., 2018).

M DETAILED RELATIONSHIP TO PRIOR WORK

Circuits, redundancy, and intervention semantics. The circuit viewpoint studies internal computations rather than only feature–output correlation (Olah et al., 2020; Elhage et al., 2021). Interpreted circuits for indirect object identification (Wang et al., 2023), automated discovery (Conmy et al., 2023), and scalable attribution-patching approximations (Kramár et al., 2024) supply important tools, but their objectives are not identical to complete recovery of a thresholded mask response. Work on the Hydra effect and self-repair (McGrath et al., 2023; Rushing & Nanda, 2024) already demonstrates compensation. Conditional Co-Ablation (CoAx) (Gong et al., 2026) explicitly recovers conditional backups. We do not claim that detecting redundancy is new. Our question is what evidence excludes further dependencies and what survives a failure of monotonicity.

Activation-patching metrics and replacement choices affect the measured causal quantity (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). Consequently all guarantees here fix a patch policy, metric, threshold, unit partition, and input distribution. Circuit overlap need not imply faithfulness (Hanna et al., 2024); our exact-family counterexample shows a further limitation even when the family overlap is perfect. A general causal abstraction (Geiger et al., 2025) is richer than the Boolean response characterized here. Formal circuit verification (Hadad et al., 2026) uses model structure to prove robustness on continuous domains and under patch perturbations. Our oracle lower bounds do not apply to unrestricted white-box verification. Statistical circuit-hypothesis tests (Shi et al., 2024) are also complementary rather than replaced by a structural cover certificate.

Hypergraph learning and teaching. An edge-detecting query is exactly the monotone intervention oracle in (2). Independent covers, edge extraction, and lower bounds are classical (Angluin & Chen, 2006). Exact monotone-DNF learning (Abasi et al., 2014), low-degree hypergraph learning (Balkanski et al., 2022), and recent several-edge search algorithms (Chen, 2026) provide stronger guarantees for different structural classes and query models. We neither replace their algorithms nor assert polynomial query complexity in general. A target-identifying transcript is a constrained teaching set (Goldman & Kearns, 1995). Our contribution is its rank/ceiling-sensitive expression and the certificate-relative adaptive analysis under the stated evaluation protocol.

Group testing has long separated query information from adaptivity and computational costs (Aldridge et al., 2026). SIGHT and Random Chemistry (Clarfeld & Eppstein, 2019) provide related search and shrinking procedures. Harmonic charging (Chvátal, 1979) and covering arrays (Colbourn, 2004) supply standard combinatorial ingredients. The global charge is useful here because discovery changes the residual problem; one fixed true-negative cover remains feasible through every phase.

Sparse interactions. Möbius-transform methods recover polynomial interactions (Kang et al., 2024; Erginbas et al., 2026). Sparsity of minimal failure sets is a different assumption. For example, the disjunction of two failure indicators is $u \vee v = u + v - uv$; combining more indicators can create many signed polynomial terms. Conversely, a sparse polynomial can describe nonmonotone cancellation. No matched polynomial-recovery comparison is evaluated here.

Monotonicity testing and exact cut optimization. Monotonicity testing asks whether a response belongs to or is far from the monotone class (Goldreich et al., 2000; Khot et al., 2018). Our error is instead relative to a specified monotone reference, constrained by transcript anchors. A response can be very close to another monotone function and very far from that reference. The sharp example makes the distinction explicit. Cube routing (Briët et al., 2012), graph-cut representations of binary energies (Kolmogorov & Zabih, 2004), and fractional programming (Dinkelbach, 1967) are established techniques. We use them to compute and bound the transcript-specific robustness coefficient, not to claim a new generic flow solver or improved property-testing complexity.

Controlled neural validation and statistics. Tracr (Lindner et al., 2023) provides controlled interpretability studies. Our transformers use standard attention (Vaswani et al., 2017), Adam (Kingma & Ba, 2015), and head dropout (Srivastava et al., 2014); exhaustive intervention evaluation supplies ground truth rather than known task rules. Reference-mass bounds and behavioral intervals use classical concentration and exact binomial inference (Hoeffding, 1963; Clopper & Pearson, 1934). Confidence sequences (Howard et al., 2021) and adaptive-data-analysis methods (Dwork et al., 2015) offer refinements; no new statistical rate is claimed.