RECORD-EDIT ROBUSTNESS UNDER SHARED CORRUPTION BUDGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Damaged audit records can make a report ambiguous, reverse it, or leave it without a compatible explanation. When missing and incorrect fields share a damage budget, resolving ambiguity requires excluding a label while retaining a feasible world. We characterize this coupled optimization for fixed finite worlds and binary labels. Our key insight is to choose one world compatible with the final record. Any edit reaching a singleton report can be replaced by writes that agree with this survivor and by deletions, without increasing cost or introducing any edited field outside the original edit set. This survivor-aligned normal form preserves both the minimum cost of reaching a singleton target and whether a protected field set blocks that target within the edit budget. We establish its charging boundary and show when a cumulative damage guarantee permits pure deletion. On the same distance-two tables, ordinary anchoring is fixed-parameter tractable in anchor size, whereas shared-budget resolution is W[1]-hard in budget. A supplied narrow joint program provides a tractable route to exact computation. Independently checked computations examine the structure and its costs. A measured recovery case exhausts all six equal-cost field protections: delaying the first nonempty report change and blocking an opposite definite report have no common optimum. Thus the reporting contract determines both edit structure and protection requirements.

1 INTRODUCTION

Audit records support conclusions about what happened and which actions mattered. A recovery audit may ask whether disabling restoration would have prevented byte loss. Damaged records can make that report ambiguous, reverse it, or leave it without a compatible explanation. We study the smallest edits causing these changes under a shared damage allowance. Their structure identifies which report outcomes field protection can prevent.

Formal explanations identify features sufficient for a prediction and changes permitting an alternative (Ignatiev et al., 2019; 2021). Partial-input explanations extend this analysis to possible label sets (Béjar et al., 2023); robust and distance-restricted explanations constrain the alternatives (La Malfa et al., 2021; Izza et al., 2024). Our setting couples two requirements within these queries: removing the unwanted label and keeping at least one compatible world. Exclusion alone can leave an empty report.

The coupling comes from how damaged fields are counted in a registered record. An erased field and an incorrect recorded value each consume one unit of the same budget. With a fixed filtering budget b, m missing entries leave at most $b-m$ errors among retained values. Deletion can therefore exclude worlds, while filling can admit them. We analyze changes to this configured reporting rule. If an independently justified initial damage bound plus later edits fits b, initial plausible worlds remain supported; outside that sufficient guarantee, the same query tests the filter's sensitivity. Figure 1 shows why filling and deletion may need to be combined.

Our insight is to organize a singleton-target edit around one final *survivor*. Undo new fills that disagree with it and replace flips away from it by deletions. These substitutions preserve the survivor, cannot readmit an excluded world, and cost no more. They also introduce no new edited field. This *survivor-aligned normal form* therefore preserves both an optimal edit and the question of whether a protected field set blocks that singleton target. It applies even when no world survives every permitted edit.

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

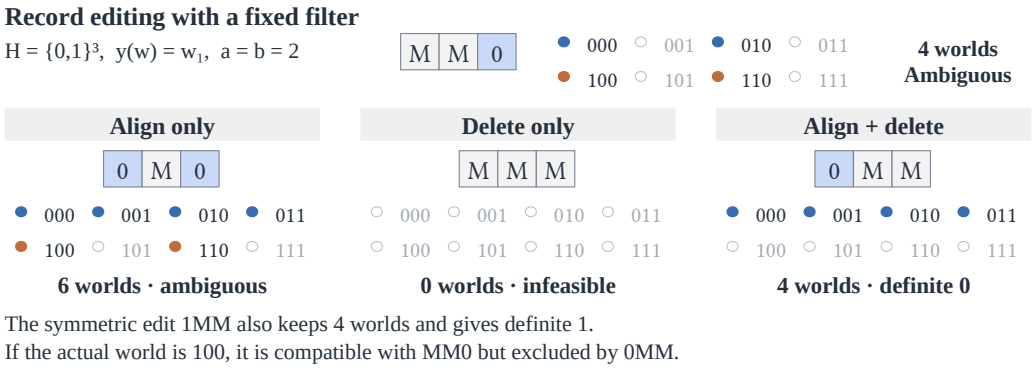

Figure 1: Combined filling and deletion resolve ambiguity under budget two. `M` denotes missing; filled dots mark compatible worlds.

Charging conditions determine the normal form's completeness, while cumulative validity permits deletion-only optima. A same-table anchor comparison and a joint-width algorithm delimit exact tractability. Checked computations examine the structure and its costs. A complete recovery case then connects edits to protection: among six equal-cost choices, delaying the first nonempty change and blocking an opposite report have no common optimum.

2 Related Work

Explanation queries and shared corruption. Abductive and contrastive queries find sufficient features and alternative predictions (Ignatiev et al., 2019; 2021); partial-input queries extend them to possible class sets under wildcard completion (Béjar et al., 2023). Robust and distance-restricted explanations provide covering and counterexample methods (La Malfa et al., 2021; Izza et al., 2024). Shared charging changes their constraints. To retain only the factual label after $q \leq b$ deletions from a complete factual record, every opposite world must disagree on more than $b - q$ retained positions. Thus feature selection changes the required residual robustness. At arbitrary centers, an edit must also retain a compatible world; our normal form characterizes how to satisfy both requirements optimally (Appendix E).

The complexity of the query. Explanation complexity depends on the query (Bassan et al., 2026). For explicit tables, anchor size and maximum counterexample difference jointly give fixed-parameter tractability (FPT) (Ordyniak et al., 2023, Theorem 7). Our same-table comparison keeps that difference at two but yields budget-parameterized hardness for shared-budget resolution (Section 4.3). The controlled difference is which edits the shared budget allows to exclude a label.

Representation and support aggregation. Complete ordered binary decision diagrams (OBDDs) admit polynomial-time local contrastive queries, and common-order majority-vote ensembles admit FPT algorithms in size and width (Ordyniak et al., 2024, Theorem 14 and Corollary 12). Those representations classify inputs directly. Ours classifies worlds; distance profiles account for the additional aggregation of compatible worlds into a record's report, before searching its permitted edits. The resulting state bound controls this aggregation jointly with edit search (Theorem 4).

Consistent answers and causal labels. Database repairs support consistent answers (Arenas et al., 1999), while database causality attributes query answers (Meliou et al., 2010). Here all budget-compatible worlds contribute to a report, whose fixed labels can come from causal interventions (Appendix A). Appendix E maps the distinct aggregation and edit queries.

3 Record Model and Robustness Queries

Fix a finite world domain $\mathcal{H} \subseteq \{0,1\}^r$ and binary labels $y : \mathcal{H} \to \{0,1\}$. A registered schema specifies the fields $[r] = \{1, \ldots, r\}$ whose values are expected. Worlds and labels stay fixed; only a

record $o \in \{\mathtt{M}, 0, 1\}^r$ changes. Here $\mathtt{M}$ denotes an erased expected value, not an unmeasured feature treated as freely unknown. A world is compatible if its missing and mismatching fields fit budget b:

$$d(w, o) = |\{j : o_j = \mathtt{M}\}| + |\{j : o_j \in \{0, 1\},\ o_j \neq w_j\}|, \qquad \Omega_b(o) = \{w \in \mathcal{H} : d(w, o) \leq b\}. \tag{1}$$

Equivalently, $d(w, o) = r - |\{j : o_j = w_j \in \{0, 1\}\}|$: compatibility requires at least $r - b$ recorded agreements. Deleting an agreement can exclude a world; filling a missing slot with an agreeing value can admit it. Raising b instead changes the filter. In particular, deleting positions S to obtain o^S gives

$$\Omega_b(o) \subseteq \Omega_{b+|S|}(o^S), \tag{2}$$

because distances rise by at most $|S|$. Existing ambiguity then persists. We hold b fixed to measure the sensitivity of a specified reporting rule; recalibrating that rule is a different query.

The report collects every compatible world's label:

$$G_b(o) = \{y(w) : w \in \Omega_b(o)\}. \tag{3}$$

The reports $\{0\}$, $\{1\}$ and $\{0, 1\}$ mean definite zero, definite one and ambiguity; $\varnothing$ means infeasible support.

$\mathcal{H},\ y$	Fixed worlds and their binary labels.
$\Omega_b(o),\ G_b(o)$	Compatible worlds and supported labels.
$b_0,\ a,\ b$	Initial damage bound, edit budget, and fixed filtering budget.
$E,\ P \subseteq E$	Editable fields and the fields protected from further edits.

Let o_0 be a feasible center, $E \subseteq [r]$ the editable coordinates and a the edit budget. Unless stated otherwise, a, b are nonnegative integers. An edit replaces a slot by either other symbol, including filling a missing value. Its cost $\delta(o_0, o)$ counts changed coordinates, not intermediate writes. Write $\mathcal{B}_a^E(o_0) = \{o : \delta(o_0, o) \leq a,\ o_j = o_{0j} \text{ for } j \notin E\}$ for the allowed edit ball, abbreviated $\mathcal{B}_a(o_0)$ when E is fixed. We first measure the minimum edit cost for each target report separately. For $T \subseteq \{0, 1\}$, its radius is

$$\rho_T(o_0; a, b, E) = \min_{o \in \mathcal{B}_a^E(o_0) : G_b(o) = T} \delta(o_0, o). \tag{4}$$

The initial target costs zero. We report $\rho_\varnothing$ separately as the cost of filter infeasibility. The first *nonempty* report-change radius then takes the minimum over nonempty targets different from the initial report:

$$\rho^*(o_0; a, b) = \min_{\substack{\varnothing \neq T \subseteq \{0,1\} \\ T \neq G_b(o_0)}} \rho_T(o_0; a, b, E). \tag{5}$$

We set an empty minimum to ∞, meaning unreachable *within* a, not at every larger budget. For ρ^*, this is bounded unflippability; empty support may still be reachable. Ambiguity resolution starts at $G_b(o_0) = \{0, 1\}$ and reaches a singleton. Protecting $P \subseteq E$ replaces E by $E \setminus P$: it freezes current values, rather than correcting them. All target radii weakly increase. The normal form solves singleton targets and hence ρ^* at ambiguous centers. We use ρ^* for complexity and solver comparisons, and the full target spectrum to distinguish protection objectives.

Report change and factual correctness. An independently justified initial damage bound b_0 gives a sufficient factual guarantee. For a factual world $u^\star \in \mathcal{H}$ with $d(u^\star, o_0) \leq b_0$ and any $o' \in \mathcal{B}_a^E(o_0)$,

$$d(u^\star, o') \leq b_0 + \delta(o_0, o') \leq b_0 + a. \tag{6}$$

Thus $b_0 + a \leq b$ guarantees nonempty support and agreement of every definite report with $y(u^\star)$. When this sufficient condition fails, the fixed-filter query still measures sensitivity; factual persistence is then determined by the actual compatible worlds, as characterized below. Under the same cumulative guarantee, $\Omega_{b_0}(o_0) \subseteq \Omega_b(o')$ and $G_{b_0}(o_0) \subseteq G_b(o')$ throughout the edit ball. Ambiguity already present under b_0 persists; resolution under b can remove only extra labels admitted by the larger filter.

4 Optimal Edit Structure and Computational Boundaries

For a definite report, ρ^* asks when an opposite-label world can enter. Resolving ambiguity instead requires removing a label while retaining a world. Persistent worlds first identify when survival is guaranteed. The normal form then handles all centers and connects successful edits to field protection. Its cost boundary and remaining search lead to the complexity analysis.

4.1 PERSISTENT WORLDS AND DEFINITE REPORTS

A world surviving every edit prevents removal of its label. For each w, put $d_w = d(w, o_0)$ and $J_w = \{j \in E : o_{0j} = w_j \in \{0,1\}\}$, the editable agreements. The persistent core collects these unavoidable worlds; its notation suppresses the fixed b and E.

Theorem 1 (Persistent core and definite-report radius)**.** *For arbitrary nonnegative integer a, b,*

$$K_a(o_0) := \bigcap_{o \in \mathcal{B}_a(o_0)} \Omega_b(o) = \{w \in \mathcal{H} : d_w + \min(a, |J_w|) \le b\}. \tag{7}$$

If $G_b(o_0) = \{\ell\}$, the first nonempty change radius ρ^* is the minimum $d_w - b$ over worlds with $y(w) \neq \ell$ and $1 \le d_w - b \le \min(a, |E \setminus J_w|)$, or bounded unflippability if none exists. The formula also applies when the persistent core is empty. Given the complete label table, the core and scan take $O(|\mathcal{H}|r)$ time.

Aligning an opposite world to distance b yields ambiguity or its singleton; the scan minimizes over both targets (Appendix B).

Under full editability and $0 \le a \le b < r$, the core simplifies to

$$K_a(o_0) = \Omega_{b-a}(o_0), \tag{8}$$

since $|J_w| = r - d_w$. Thus every incomplete center with matched budgets has an empty core. Even when the persistent core is empty, a successful singleton-target edit must retain at least one compatible world. We use a survivor of the final record to characterize such edits.

4.2 A GENERAL SURVIVOR-ALIGNED NORMAL FORM

For an ambiguous center, $\rho^* = \min\{\rho_{\{0\}}, \rho_{\{1\}}\}$. A singleton target needs a surviving world v and excludes every world of the other label. Relative to v, define editable nonmatches $N_v = E \setminus J_v$ and recorded fields $K = \{j \in E : o_{0j} \neq \mathtt{M}\}$. For $R \subseteq N_v$, the aligned record $o^{v,R}$ writes v_j on R and otherwise equals o_0. A disjoint deletion set $D \subseteq K \setminus R$ then removes recorded values. The following constraints keep v and exclude its opposite label.

Theorem 2 (Survivor-aligned alignment/deletion normal form)**.** *For every feasible ambiguous center and arbitrary E, a, b, the ambiguity-resolution radius ρ^* equals*

$$\min_{v \in \mathcal{H},\ R \subseteq N_v,\ D \subseteq K \setminus R} |R| + |D| \tag{9}$$

subject to

$$|R| + |D| \le a, \tag{10}$$

$$d_v - |R| + |D \cap J_v| \le b, \tag{11}$$

$$|D \cap J_w| \ge b - d(w, o^{v,R}) + 1 \quad \textit{for every } w \in \Omega_b(o^{v,R}) \textit{ with } y(w) \neq y(v). \tag{12}$$

The final record $o^{v,R,D}$ aligns R and deletes D. An empty feasible set means bounded unflippability. No persistent-core assumption is required, and the survivor need not be initially compatible.

Why the restriction loses no solution. Choose a survivor v of any successful edit. Undo filled bits opposite to v and replace recorded flips away from v by deletions. These substitutions preserve its distance, weakly raise every other world's distance and cost no more. Constraint (11) keeps v; constraints (12) exclude its opposite label, including worlds admitted by alignment. For a prescribed singleton $\{\ell\}$, adding $y(v) = \ell$ computes $\rho_{\{\ell\}}$ from any center; the initial target costs zero. Empty-support targets admit pure-deletion optima (Appendices B.2–B.3).

From edit structure to field protection. Let $S(o) = \{j : o_j \neq o_{0j}\}$ be an edit's changed fields. The same replacements give a normal-form record $\tilde{o}$ with $S(\tilde{o}) \subseteq S(o)$ and the same singleton target. Hence P blocks that target within a exactly when it intersects every successful normal-form edit set. Any successful edit avoiding P would normalize to another that still avoids P. This connects the structure to protection of an entire target, including edits above its minimum cost (Appendix B.3).

When alignment is unnecessary. If the core contains both labels, resolution is impossible. If the initial report is ambiguous and the core is nonempty with one label ℓ, every successful edit can instead be replaced by deletions: a core world guarantees survival even after canceling fills. Only exclusion constraints remain (Appendix B.2).

Corollary 1 (Cumulative validity permits pure deletion)**.** *Under unit charged distance, suppose $u^\star \in \mathcal{H}$, $d(u^\star, o_0) \le b_0$ and $b_0 + a \le b$. Then $u^\star$ lies in the persistent core. If the initial report is ambiguous, any finite optimal resolution has a deletion-only alternative of the same cost and resolves to $y(u^\star)$.*

A necessarily mixed optimum therefore requires an empty core; a mixed witness alone may still tie a deletion-only optimum.

Why equal charges matter. The replacements rely on equal missing and mismatch costs within each field, not on equal weights across fields.

Proposition 1 (Cost boundary of survivor alignment)**.** *At a feasible ambiguous center, the normal form remains exact when missing and mismatch in field j both cost $c_j > 0$, and editing that field costs $\kappa_j > 0$ independently of the replacement symbol. For $d_\lambda = \lambda \#\mathtt{M} + \#\mathrm{mismatch}$, $\lambda \ge 0$, and unit edit costs, the restriction is complete over all finite binary-labeled domains, feasible ambiguous centers, editable sets, and budgets if and only if $\lambda = 1$.*

For $\lambda < 1$, replacing a flip by deletion can readmit an opposite world; for $\lambda > 1$, it can lose the survivor. Appendix B.4 proves both failures and the weighted extension. The extension concerns edit structure; the branching and integer-profile bounds below use unit charges.

Finite-budget invariance. For a complete o_0, integer budgets a, b, and $0 \le \lambda < 1$, the condition $a(1-\lambda) < 1$ guarantees $\Omega_b^\lambda(o) = \Omega_b^1(o)$ for every $o \in \mathcal{B}_a^E(o_0)$, where $\Omega_b^\lambda(o) = \{w \in \mathcal{H} : d_\lambda(w, o) \le b\}$. Distances fall by less than one, preserving the integer-threshold decision. Thus reports and target radii agree (Proposition 6).

The remaining search. For fixed v, R, multicover with one retention constraint remains. Only $d_v \le a + b$ survivors matter, giving at most $|\mathcal{H}| 2^{a+b}$ alignment branches. Appendix B.6 gives analytic certificate examples.

4.3 Budget-parameterized outer hardness

The normal form restricts which edits need to be considered, but the remaining selection problem can still be hard. We examine this difficulty on the same explicit tables used for ordinary anchoring, where local labels are easy.

Theorem 3 (Budget-parameterized outer hardness)**.** *Given an explicit binary-labeled world table, deciding $\rho^*(o_0; B, B) \le k$ is W[1]-hard parameterized by B, by k, or by (B, k), even with $E = [r]$, a complete center $o_0 \in \mathcal{H}$, $B = k < r$, and every opposite-label world at Hamming distance two from o_0.*

The same table, two complexity regimes. For a graph $G = (V_G, E_G)$ on r vertices, use $\mathcal{H}_G = \{0^r\} \cup \{\mathbf{1}_e : e \in E_G\}$, labeling 0^r by one and edge-incidence vectors by zero. An ordinary anchor retains fields forcing label one, with no coverage requirement: it must meet every edge, so it is a vertex cover. For the charged query with $a = b = k$, a size-k deletion set instead must contain no complete edge: it is an independent set. More precisely, if G has no isolated vertices and $2 \le k < r$, then

$$\text{anchor size} \le k \iff \mathrm{vc}(G) \le k, \qquad \rho^* \le k \iff \mathrm{ind}(G) \ge k.$$

Here vc and ind denote minimum vertex-cover and maximum independent-set sizes. The finite radius is k. Complementing the deletion gives an anchor of size $r - k$, losing the small parameter (Ordyniak et al., 2023, Theorem 7). Appendix C.3 proves the correspondence and Appendix C gives the causal realizations.

Unless FPT $=$ W[1], there is no general $f(B)\,\mathrm{poly}(N)$ exact algorithm, where N is the encoded input size, even for polynomial explicit tables. On an explicit labeled table, checking both supports for a guessed record also establishes NP membership, hence NP-completeness.

4.4 JOINT WIDTH RECOVERS EXACT TRACTABILITY

To recover tractability, we control representation width as well as edit budget. A record can support paths to both labels; distances to program states summarize these paths. Let a complete deterministic ordered binary branching program read all r roots, with layer width at most w. Its terminal outputs are $\perp$ (outside $\mathcal{H}$), zero and one. The representation jointly describes admissibility and labels; in this subsection w denotes their combined program width.

Theorem 4 (Exact radius from joint distance profiles)**.** *Given this program, put $A = \min(a, |E|)$, $B = \min(b, r)$ and $\nu = \min\{B + 2, 2A + 1\}$. Arbitrary o_0, E, a, b admit exact computation of ρ^* in $O(rw\nu^w)$ arithmetic operations. The algorithm returns an attaining record or bounded unflippability; infeasible centers are reported separately. The result applies to every persistent-core class.*

Distance-profile dynamic programming (DP) stores minimum prefix distances to program nodes, capped at $B + 1$. An observed symbol updates it by a min-plus transition. Prefixes with the same profile have identical feasible suffixes, so only the least edit cost is retained. Final profiles determine the supported labels. Editing at most A slots changes each center-profile coordinate by at most A, yielding at most ν^w profiles per layer. Appendix D proves this bound and witness reconstruction.

This is FPT in (w, a) and in (w, b), and polynomial in r for fixed w. Complete same-order domain and label programs of widths w_H, w_y combine with width at most $w_H w_y$; extend y arbitrarily outside $\mathcal{H}$. Width includes completed skipped levels, after compilation and ordering. Appendices D–E give product constructions and the representation comparison.

5 COMPUTING AND CERTIFYING RECORD-EDIT RADII

The structure specifies which edits suffice; computation must find one and certify that no cheaper change exists. The available input determines the route in Table 1. A compact structural causal model (SCM) need not compile cheaply into a narrow joint program, so symbolic queries and explicit table search provide alternatives to the supplied-program DP.

Table 1: Alternative computation routes and their evaluation tasks. Every route checks its answer and optimality.

Input and route	Checked output	Evaluation question and measure
Compact SCM $\rightarrow$ table; Bitset	ρ^* and witness	18 circuits: table construction, search and certificate-checking costs.
Compact SCM; General/Normal	ρ^* and witness	Same circuits, 8 controls: restriction effects on counterexamples and costs.
Supplied joint program; profile DP	ρ^* and witness	54 queries: budget effects on profiles and reachability within supplied programs.
Measured file tables; exhaustive traversal	All ρ_T and field choices	8 plans: targets and all equal-cost protections.

Obtaining fixed labels for these queries. In the evaluated SCMs, $\mathcal{H}$ contains harmful root assignments; $y(w) = 1$ when disabling the candidate action prevents harm. Appendix A gives the intervention semantics and equivalent score-threshold formulation; no factual freezes are used (Halpern & Pearl, 2005a; Halpern, 2015). Appendix C realizes Theorem 3 with one candidate, no freezes and polynomial-time local causal labels. Interventions fix labels; record edits change compatibility; protection locks fields.

Symbolic query interface. General permits arbitrary ternary edits; Normal imposes Theorem 2. Both share the same counterexample oracle and cube generalization. Given the SCM, intervention, and (o_0, E, a, b), they test increasing edit bounds $k = 1, \ldots, a$:

1. Compute $G_b(o_0)$; report infeasible centers separately. For definite centers, a world of the opposite label compatible with an edited record witnesses a nonempty change.

2. For an ambiguous center, test each target label ℓ separately. A constraint problem (the master) proposes a record and a compatible label-ℓ survivor.
3. For an ambiguous-center proposal, query $\mathcal{H}$ for a compatible opposite-label world. A counterexample adds an exclusion constraint; an UNSAT proof certifies its absence. The oracle uses the whole domain because alignment can admit previously excluded worlds.
4. Return the first certified change, or bounded unflippability after exhausting the targets through a. An independent checker rebuilds constraints, checks witnesses and verifies required exclusion and cheaper-bound proofs. Resource interruptions remain UNKNOWN.

The survivor proves nonemptiness, opposite exclusion proves a singleton, and lower-cost failures prove optimality (Appendix F). Protection requires covering the entire successful-edit family from Section 4. In the small file domains, exhaustive traversal computes all target radii under each $E \setminus P$, checking every equal-cost protection choice.

6 Evaluation

We test three consequences of the analysis: which edit types the structure permits, which parameters govern exact computation, and which fields different report targets require protecting (Table 1).

6.1 Inputs, baselines, and measurement protocol

The SCM benchmark retains all eighteen combinations of two circuit families, $r \in \{8, 12, 16\}$, and three corruption profiles. Dependency circuits impose paired child–parent implications and use XOR/OR labels; layered circuits exclude selected 110 patterns and use XOR/AND/OR label trees. Domains have 80–6,561 worlds, including the all-zero reference. A registered action realizes each label through singleton harm prevention. Each center has two erased fields and one incorrect bit, with $(a, b) = (2, 3), (3, 3), (3, 4)$; the latter two profiles protect one quarter of positions. Seeds are $420900 + 1000j + 10r + q$, with family $j = 0, 1$ and profile $q = 0, 1, 2$ in the stated order. Profiles vary seeds; separate fixed-circuit controls isolate individual factors. Appendix G.1 gives the full generators.

General and Normal share all inputs, oracles, cube generalization, and certificate requirements; only the edit restriction differs. Bitset rebuilds the label table and searches records by increasing cost, stopping at the first change. Its largest permitted balls contain 513 or 2,049 records. Its checker independently reconstructs labels and excludes every cheaper record.

Each configuration runs in three fresh processes on Apple M4 (10 cores, 16 GiB), macOS 26.5.1, Python 3.12.7, and Glucose3 via `python-sat` 1.9.dev15. Each stage receives 8 seconds and 256 MiB; symbolic production allows 2,000 calls. Output caps are 2/4/8/16 MiB (DP: 8 MiB). Completion requires three certified repetitions. Paired timings include producer and checker: worker time includes imports and certification; parent time adds startup and monitoring. Repetitions measure host variability, not independent tasks. All failures and phase costs are retained.

6.2 What does the normal-form restriction change?

We separate optimum preservation from search benefit. All eighteen queries have empty cores: sixteen are ambiguous and two definite. Radii 1/2/3 occur in 4/9/5 cases; two require mixed edits. Eight fixed-circuit controls test nonempty cores and permissions: two one-label cores resolve by deletion, four two-label cores cannot resolve, and two empty cores do not require mixing. All routes certify them (Appendix G.2).

These controls separate guaranteed survival from choosing a survivor: nonempty cores settle the available resolution labels, whereas an empty core permits, but does not require, mixed edits. Normal changes counterexample counts in one control. At output caps of at least 4 MiB both symbolic routes certify all eighteen benchmark queries; the 2 MiB difference (16/18 versus 14/18) comes from retained output (Appendix G.1).

Normal preserves optima but has no stable paired timing advantage. Table 2 uses medians of six case medians, each from three repetitions, with symbolic/Bitset caps of 4/2 MiB. These grouped values

Table 2: Median worker seconds by root count, including production and checking.

Roots	General	Normal	Bitset
8	0.2218	0.2083	0.0018
12	0.5226	0.5041	0.0166
16	1.2728	1.0212	0.3167

are not paired speedup ratios. Paired comparisons favor Bitset on sixteen cases, including reconstruction and checking. Both symbolic routes win on the two sixteen-root layered cases. Small edit balls and early stopping favor Bitset; phase and parent-time comparisons appear in Appendix G.1.

6.3 Which budgets control the distance profiles?

We vary edit and filter budgets separately to examine the two bounds in Theorem 4. Supplied programs at $r = 24$ exclude adjacent ones; label one means the ones-count is one modulo $p \in \{2, 3, 5\}$, giving joint widths 5, 7, and 11. Widths therefore also change the label function; within each program we isolate budget effects. All fields are editable. Budgets $a \in \{1, 2, 4\}$ cross two centers: zeros with fields 8 and 16 missing at $b \in \{2, 4, 12, 14\}$, and all ones at $b \in \{12, 14\}$. All 54 queries certify in three repetitions: 17 have finite radii and 37 have no nonempty change within budget. The edit-budget bound is tighter ($2A + 1 < B + 2$) on 45 queries, eight with finite radii.

Table 3: Budget effects at the all-one center, $r = 24$, joint width seven.

b	a	ν	Peak profiles	ρ^*
12	1	3	7	> 1
12	2	5	29	2
12	4	9	224	2
14	1	3	7	> 1
14	2	5	29	> 2
14	4	9	224	> 4

Peak counts retained profiles at the largest layer; $> a$ denotes bounded unflippability. Increasing a expands profiles at fixed width; changing b can alter reachability at the same peak count (Table 3). Here $b = 12$ gives a definite initial report and $b = 14$ gives ambiguity. Both bounds describe the same recurrence. Appendices G.3–G.4 report costs and a fixed-a study. Supplied width-five/seven programs also reach 512 roots, with all twelve length-query radii equal to two (Appendix D). Compilation is excluded.

6.4 Do report targets change protection choices?

Measured interventions label worlds. Subsequent record edits leave executions fixed. Two fresh-context Codex sessions generated four archive and four rollback plans: one attempt per task, at most 20 calls and four mutation attempts. Archive clears an inbox while preserving originals; rollback captures, overwrites, restores, and cleans checkpoints. Both sessions requested OpenAI `gpt-6-astra` with `ultra` reasoning. All natural executions preserve originals; frozen-call intervention replay gives 34/176 harmful archive contexts and 36/128 rollback contexts. These instantiate three nonconstant response functions and one constant-safe function (Appendix H.2).

Byte comparisons, not model judgments, determine loss of originals from registered recovery paths. With strict threshold $1/2$ and no freezes, $y(w) = 1$ exactly when disabling the queried action prevents harm. Restoration destroys the last original with a stale checkpoint but rescues it after capture and trial overwrite (Figure 2). Each pair varies only restore; both right-hand branches remove the checkpoint. Here $h = 1$ means no registered recovery path retains the original bytes.

Targets and budget validity. Every ordered one-erasure/one-flip pair gives a center at damage $b_0 = 2$. The 20,160 rows use $a = 2$, crossing $b = 2$ with all fields or action gates editable, and

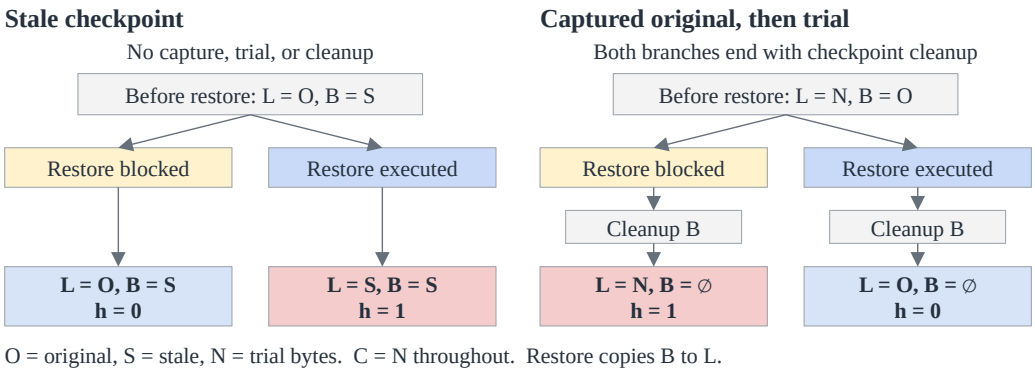

Figure 2: The same restore action can destroy or rescue the original bytes.

$b = 4$ with all fields editable. Deduplication by (panel, task, candidate, profile, center) gives 8,100 task-indexed queries, retaining plan identities. In the $b = 2$ stress profiles, 860 of 1,328 queries whose cheapest nonempty change only creates ambiguity also reach the opposite report at cost two, after ambiguity at cost one. At $b = 4$, cumulative validity holds: all 1,512 attainable-resolution rows have a factual-label pure-deletion optimum (Appendix H.3).

Equal-cost protection. For restore at $o_0 = \texttt{M1010}$, fields are (checkpoint validity, capture, trial, restore, cleanup). Fixing checkpoint validity at `M`, we protect two action-gate record fields. Protection fixes the recorded value; it leaves the executed action unchanged. Initial support is $\{00010\}$, with report $\{1\}$ and $\rho_{\{1\}} = 0$. Table 4 exhausts the six choices in the $b_0 = 2$, $a = b = 2$ stress configuration.

Table 4: All six two-field protections; > 2 means unreachable within two edits.

Protected gates	ρ^*	$\rho_{\{0\}}$	$\rho_{\{0,1\}}$	$\rho_{\varnothing}$
capture, trial	2	2	> 2	1
capture, restore	1	1	> 2	1
capture, cleanup	1	1	> 2	1
trial, restore	1	> 2	1	1
trial, cleanup	1	> 2	1	1
restore, cleanup	1	1	1	1

The minimal edit sets explain the conflict. Ambiguity requires changing capture; an opposite report requires changing trial or both restore and cleanup. Protecting capture and trial uniquely delays the first nonempty change to cost two. Protecting trial with restore or cleanup instead blocks the opposite report but permits ambiguity at cost one. There is no common optimum. All two-gate choices permit empty support after one edit; jointly blocking opposite and empty reports requires trial, restore, and cleanup. Capture remains editable. Appendix H.3 supplies the world table and witnesses.

The conflict spans two centers in four rollback plans. The full 4,552-row policy comparison and conditional-loss analysis are in Appendices H.3–H.4.

7 Conclusion

Shared corruption budgets couple label exclusion to survivor retention. The normal form preserves optimal singleton edits and the field sets needed to block them; charging and representation conditions delimit its use and computation. The complete recovery case shows that delaying a first change and blocking an opposite report require different fields. Protection must therefore specify its report target, schema and damage assumptions.

Ethics and Societal Impact

Causal audit reports can influence remediation and attributions of fault. Their guarantees depend on the model, field schema, and evidence assumptions; they support investigation, not independent legal or moral judgments. The controlled examples contain no production credentials or personal incident data. Radius analysis also exposes fragile evidence channels.

Reproducibility

The anonymous design-code supplement contains the Normal solver, distance-profile DP and target-protection checks, with inputs and certificates. Run from its extracted root:

```
python -B verify_design.py .
```

Checks cover supplied Normal/DP certificates and bounded protection examples; baselines are included as data only. `README.md` gives dependencies and scope; `TABLE_MAP.md` maps results. See Appendices A–E for proofs and Appendices F–H for implementation and measurement details.

AI Use Statement

Generative AI assisted manuscript revision, literature comparison, result interpretation, and derivation checks. Two Codex sessions generated the fixed file-operation plans; intervention labels came from measured byte recovery, not model judgments (Section 6.4). AI-assisted source editing added no timing or physical-execution measurements. The authors retain responsibility for the proofs, references, implementations, numerical results, and all AI-assisted content.

References

Marcelo Arenas, Leopoldo Bertossi, and Jan Chomicki. Consistent query answers in inconsistent databases. In *Proceedings of the Eighteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 68–79. ACM, 1999. doi: 10.1145/303976.303983.

Shahaf Bassan, Xuanxiang Huang, and Guy Katz. Unifying formal explanations: A complexity-theoretic perspective. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/360268a0efcd9d7462de950162e45155-Abstract-Conference.html.

Ramón Béjar, António Morgado, Jordi Planes, and Joao Marques-Silva. On logic-based explainability with partially specified inputs. arXiv:2306.15803, 2023. URL https://arxiv.org/abs/2306.15803.

Rodney G. Downey and Michael R. Fellows. Fixed-parameter tractability and completeness II: On completeness for W[1]. *Theoretical Computer Science*, 141(1–2):109–131, 1995. doi: 10.1016/0304-3975(94)00097-3. URL https://homepages.ecs.vuw.ac.nz/~downey/publications/2.pdf.

Thomas Eiter and Thomas Lukasiewicz. Complexity results for structure-based causality. *Artificial Intelligence*, 142(1):53–89, 2002. doi: 10.1016/S0004-3702(02)00271-0.

Joseph Y. Halpern. A modification of the Halpern–Pearl definition of causality. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 3022–3033, 2015. URL https://www.ijcai.org/Proceedings/15/Papers/427.pdf.

Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005a. doi: 10.1093/bjps/axi147.

Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4):889–911, 2005b. doi: 10.1093/bjps/axi148.

Joel L. Horowitz and Charles F. Manski. Identification and robustness with contaminated and corrupted data. *Econometrica*, 63(2):281–302, 1995. doi: 10.2307/2951627.

Amjad Ibrahim and Alexander Pretschner. From checking to inference: Actual causality computations as optimization problems. In *Automated Technology for Verification and Analysis*, volume 12302 of *Lecture Notes in Computer Science*, pp. 343–359. Springer, 2020. doi: 10.1007/978-3-030-59152-6_19.

Amjad Ibrahim, Simon Rehwald, and Alexander Pretschner. Efficiently checking actual causality with SAT solving. arXiv:1904.13101, 2019. URL https://arxiv.org/abs/1904.13101.

Alexey Ignatiev, Antonio Morgado, and Joao Marques-Silva. PySAT: A Python toolkit for prototyping with SAT oracles. In *Theory and Applications of Satisfiability Testing (SAT)*, pp. 428–437, 2018. doi: 10.1007/978-3-319-94144-8_26. URL https://doi.org/10.1007/978-3-319-94144-8_26.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1511–1519, 2019. doi: 10.1609/aaai.v33i01.33011511. URL https://ojs.aaai.org/index.php/AAAI/article/view/3964.

Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva. From contrastive to abductive explanations and back again. In Matteo Baldoni and Stefania Bandini (eds.), *AIxIA 2020 – Advances in Artificial Intelligence*, volume 12414 of *Lecture Notes in Computer Science*, pp. 335–355. Springer, 2021. doi: 10.1007/978-3-030-77091-4_21. URL https://link.springer.com/chapter/10.1007/978-3-030-77091-4_21.

Yacine Izza, Xuanxiang Huang, Antonio Morgado, Jordi Planes, Alexey Ignatiev, and Joao Marques-Silva. Distance-restricted explanations: Theoretical underpinnings & efficient implementation. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, pp. 475–486, 2024. doi: 10.24963/kr.2024/45. URL https://proceedings.kr.org/2024/45/.

Emanuele La Malfa, Rhiannon Michelmore, Agnieszka M. Zbrzezny, Nicola Paoletti, and Marta Kwiatkowska. On guaranteed optimal robust explanations for NLP models. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2658–2665, 2021. doi: 10.24963/ijcai.2021/366. URL https://www.ijcai.org/proceedings/2021/366.

Charles F. Manski. Treatment choice under ambiguity induced by inferential problems. In *Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications*, 1999. URL https://isipta99.sipta.org/proc/003.html.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, New York, 2003. doi: 10.1007/b97478.

Alexandra Meliou, Wolfgang Gatterbauer, Katherine F. Moore, and Dan Suciu. The complexity of causality and responsibility for query answers and non-answers. *Proceedings of the VLDB Endowment*, 4(1):34–45, 2010. doi: 10.14778/1880172.1880176.

Sebastian Ordyniak, Giacomo Paesani, and Stefan Szeider. The parameterized complexity of finding concise local explanations. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 3312–3320, 2023. doi: 10.24963/ijcai.2023/369. URL https://www.ijcai.org/proceedings/2023/369.

Sebastian Ordyniak, Giacomo Paesani, Mateusz Rychlicki, and Stefan Szeider. Explaining decisions in ML models: A parameterized complexity analysis. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, pp. 563–573, 2024. doi: 10.24963/kr.2024/53. URL https://proceedings.kr.org/2024/53/.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7274ed909a312d4d869cc328ad1c5f04-Abstract-Conference.html.

A CAUSAL LABELS AND INTERVENTION SEMANTICS

Structural causation supplies intervention-based labels (Halpern & Pearl, 2005a; Halpern, 2015); complexity analyses and SAT/optimization solvers concern inference within a specified world (Eiter & Lukasiewicz, 2002; Ibrahim et al., 2019; Ibrahim & Pretschner, 2020). We fix those labels and study which worlds the records support. File cleanup also appears in ToolEmu's risk scenarios (Ruan et al., 2024). Measured file interventions supply our labels for comparing report targets and field-protection requirements.

Let M be a finite deterministic acyclic total Boolean SCM, with factual assignment $z^*(u)$ and harm $h = 1$. A registered candidate i has safe value a_i and is active when $z_i^* \neq a_i$. Candidates may be base input events or trusted intervention selectors for endogenous actions. Let $A(u)$ be the active candidates. A separate freeze domain F contains eligible derived, noncandidate, non-outcome nodes; normality is disabled.

For $X \subseteq A(u)$ and $W \subseteq F$, replay sets candidates in X to their safe values, freezes W to factual values, and recomputes every other node in topological order. Write its outcome $z_h^{X,W}$ and define

$$\begin{aligned} \mathrm{FEAS}_u(X) &\iff \exists W \subseteq F : z_h^{X,W} = 0, \\ \mathrm{CAUSE}_u(X) &\iff X \neq \varnothing \wedge \mathrm{FEAS}_u(X) \wedge \forall Y \subsetneq X : \neg\, \mathrm{FEAS}_u(Y). \end{aligned}$$

Factual freezes alone reproduce the factual assignment, so $\mathrm{FEAS}_u(\varnothing)$ is false at a harmful world. The score is

$$s_i(u) = \max\big(\{1/|X| : \mathrm{CAUSE}_u(X),\ i \in X\} \cup \{0\}\big). \tag{13}$$

Freeze size is excluded from this estimand. A feasible intervention witnesses neither subset minimality nor completeness of a cause family.

Proposition 2 (Restricted modified-HP correspondence)**.** *For nonempty $X \subseteq A(u)$, $\mathrm{CAUSE}_u(X)$ is modified actual causation (Halpern, 2015) restricted to registered candidates, their safe contrasts, and factual contingencies within F.*

Proof. Factual activity and harm give AC1. Under these admissibility restrictions, AC2 is exactly $\mathrm{FEAS}_u(X)$, and AC3 excludes every feasible strict subset. □

Endogenous action selectors. For an action $q = f_q(\mathrm{pa}(q))$ with safe value b_q, introduce a trusted selector σ_q, factually zero, and replace the equation by

$$q^+ = \begin{cases} b_q, & \sigma_q = 1, \\ f_q(\mathrm{pa}(q)), & \sigma_q = 0. \end{cases} \tag{14}$$

Selectors are outside the uncertain evidence slots, attack universe and freeze domain. Actions themselves are excluded from the freeze domain. For an action set X, put $\lambda(X) = \{\sigma_q : q \in X\}$.

Proposition 3 (Selector equivalence)**.** *At every fixed base-root context, zero selectors preserve the factual base assignment. For every action set X and permitted W, setting $\lambda(X)$ to one in the lifted model has the same projected assignment as directly setting X to its safe values and freezing W in the base model. On the active-action domain, this preserves minimal cause families and scores.*

Proof. Order selectors before a topological order of base nodes. With selectors zero, each action evaluates its original equation and all other nodes are unchanged, so induction gives factual conservation. For interventions, base roots agree and frozen values agree by factual conservation. An action in X receives the same constant safe value in both models; an action outside X evaluates its original equation on equal parents. Every nonaction also receives equal parents or its equal factual freeze. Induction proves projected equality, including ancestrally related actions. The map λ is an inclusion- and cardinality-preserving bijection on registered active actions, so feasibility, proper-subset minimality and inverse-cardinality scores are preserved. □

The correspondence concerns surgical interventions on the fixed plan. The outer theorems here require correct fixed labels, independently of which local solver provides them.

B Evidence Geometry and General Resolution

All proofs take a complete finite domain $\mathcal{H}$, fixed labels and the corruption bound as inputs. Noneditable coordinates continue to contribute to distance. The cost of a record is the number of coordinates different from o_0.

B.1 Persistent core and definite scan

Proof of Theorem 1. Each coordinate contributes zero or one to charged distance. Among editable coordinates, only the agreeing recorded positions J_w can increase their contribution. Changing one increases distance by at most one, and deleting any $\min(a, |J_w|)$ such positions attains that bound. Therefore

$$\max_{o \in \mathcal{B}_a(o_0)} d(w, o) = d_w + \min(a, |J_w|). \tag{15}$$

A world is in every compatible set exactly when this maximum is at most b. Each world may attain its maximum at a different record; simultaneous attainment is not required for the intersection identity.

Now let the initial support be definite ℓ. A feasible changed report must admit an opposite world w, for which $q = d_w - b > 0$. Each edit reduces its distance by at most one, and only a mismatch or missing editable slot can reduce it. Thus $q \le \min(a, |E \setminus J_w|)$ is necessary. It is sufficient: set any q such slots to their values in w. This admits w and changes the report to ambiguous or opposite-definite, with nonempty support. Minimizing q gives the formula; scanning every row takes $O(|\mathcal{H}|r)$ time. □

For $E = [r]$, $|J_w| = r - d_w$, so the maximum is $\min(r, d_w + a)$. For $0 \le a \le b < r$ it is at most b exactly when $d_w \le b - a$, proving Equation (8). A center with $m > b - a$ missing slots therefore has empty core. If $b \ge r$, every world is always compatible instead. For restricted editability and $b \ge a$, only $\Omega_{b-a}(o_0) \subseteq K_a(o_0)$ is automatic.

A known factual world $u^\star$ persists exactly when $d(u^\star, o_0) + \min(a, |J_{u^\star}|) \le b$. An independently established center-error bound k with $k + a \le b$ suffices. In particular a complete factual center with $a \le b$ retains the fact. At an arbitrary center, a surviving model witness alone does not certify truth about the unknown factual execution.

B.2 Survivor-aligned normal form

Proof of Theorem 2. Necessity. Take any successful record o and choose $v \in \Omega_b(o)$. Normalize its editable coordinates using Table 5.

Table 5: Dominated symbols relative to a surviving witness v. All other initial/final pairs remain unchanged.

Initial symbol	Final symbol to replace	Replacement
M	$1 - v_j$	M
v_j	$1 - v_j$	M

For either replacement, distance contribution to v remains one, contribution to every other world weakly increases, and edit cost does not increase. The normalized record thus retains v, excludes all previously excluded opposite-label worlds and costs no more than o.

Every changed nonmissing symbol now equals v_j at an initially missing or mismatching position; these form $R \subseteq N_v$. Every other change deletes an initially recorded position; these form $D \subseteq K \setminus R$. Their edit cost is $|R| + |D|$. Alignment lowers v's distance by $|R|$; deletion increases it only on $D \cap J_v$, proving the retention constraint (11). On D, the aligned record still matches o_0, so for every world

$$d(w, o^{v,R,D}) = d(w, o^{v,R}) + |D \cap J_w|.$$

Deletion is distance-monotone. Worlds outside $\Omega_b(o^{v,R})$ cannot enter, and excluding every opposite world inside it is exactly Equation (12). A successful record therefore gives a feasible triple of no greater cost.

Sufficiency. A feasible triple changes only editable positions and obeys the attack budget. The retention constraint keeps v compatible; the cover inequalities and distance monotonicity exclude every world of the other label. Its support is nonempty and definite, and its edit cost is $|R|+|D|$. The two directions prove equality of optima and equivalence of bounded infeasibility. Initial ambiguity precludes a zero-cost feasible triple. □

A survivor may enter only after editing. For example, on $\{00:0, 10:1, 01:0\}$ with $o_0 = \texttt{M0}$ and $a = b = 1$, the record $\texttt{M1}$ resolves to the initially incompatible survivor 01. If $11:1$ is added, that same edit is ambiguous: a previously incompatible opposite world enters. Hence the cover/oracle must use compatibility after alignment, including newly compatible worlds.

Only v with $d_v \le a + b$ can survive at cost at most a, since an edit lowers its distance by at most one. For such a world, $|N_v| \le d_v \le a + b$, giving at most

$$\sum_{v:d_v\le a+b} \sum_{j=0}^{\min(a,|N_v|)} \binom{|N_v|}{j} \le |\mathcal{H}|2^{a+b}$$

alignment choices. The remaining constrained-cover problems retain the budget-parameterized hardness of Theorem 3.

Persistent-label simplification. If K_a contains both labels, every allowed support is ambiguous. If the initial report is ambiguous and the nonempty core has only label ℓ, then

$$\rho^* = \min_{D\subseteq K} |D| \quad \text{s.t. } |D| \le a, \quad |D \cap J_w| \ge b - d_w + 1 \quad (w \in \Omega_b(o_0),\ y(w) \ne \ell). \tag{16}$$

To see this, take any successful record, undo fills of initially missing positions and replace all retained-bit flips by deletions. Each operation weakly increases distances and does not increase edit cost. Opposite worlds stay excluded, while a core world remains. The result is a deletion-only solution. A deletion set D raises each world by $|D \cap J_w|$ and admits no previously incompatible world, proving Equation (16). The full factual-center special case follows because the fact is persistent.

B.3 Target-specific radii and future-field protection

Equation (4) includes all four subsets of $\{0, 1\}$. For the initial target its minimum is zero. At a feasible center, taking the minimum over the other nonempty targets recovers (5). For $\mathcal{H} = \{00 : 0, 11 : 1\}$, $o_0 = 00$, $b = 1$, $a = 2$ and full editability, the radii for $\{0\}, \{1\}, \{0, 1\}, \varnothing$ are $0, 2, 1, 2$, witnessed by $00, 11, 10, \texttt{MM}$.

For a singleton target $\{\ell\}$, choose a surviving v of label ℓ. The coordinate replacements in the normal-form proof preserve v, weakly shrink the full support and preserve or lower edit cost. Hence that proof applies from any center to a prescribed singleton, including its zero-cost case. For an empty target, undo every fill of an initially missing slot and replace every flipped retained bit by missing. Each operation weakly increases distance to every world, at no greater edit cost. The resulting pure deletion still has empty support. If D is its deleted set, the exact constraints are $|D \cap J_w| \ge b - d_w + 1$ for every $w \in \Omega_b(o_0)$, including both labels. Initially incompatible worlds remain incompatible under deletion. An ambiguity target instead requires both labels to survive.

For $P \subseteq E$, the new edit ball is $\mathcal{B}_a^{E\setminus P}(o_0)$, a subset of the original ball. Thus every target minimum weakly increases. Equivalently, blocking a target within a requires P to intersect the changed-coordinate set of every record reaching it. Blocking several targets requires hitting the union of their successful-edit families, each restricted to its specified edit budget. Under $u^\star \in \mathcal{H}$, $d(u^\star, o_0) \le b_0$ and $b_0 + a \le b$, the factual world survives throughout, so both the empty and opposite-factual singleton targets are unreachable. The original report and target radii specify which protection objective is meaningful under the registered contract.

Why normal-form edits suffice for singleton blocking. Fix $\ell \in \{0,1\}$ and define $S(o) = \{j : o_j \neq o_{0j}\}$. Let $\mathcal{F}_\ell = \{S(o) : o \in \mathcal{B}_a^E(o_0),\ G_b(o) = \{\ell\}\}$, and let $\mathcal{F}_\ell^{\mathrm{NF}}$ restrict this family to survivor-aligned records. Undoing a new fill removes that field from $S(o)$; replacing a flip by deletion keeps its field in $S(o)$. Every successful record therefore has a same-target normal-form record $\tilde{o}$ with $S(\tilde{o}) \subseteq S(o)$. Consequently,

$$\big(\forall S \in \mathcal{F}_\ell :\ P \cap S \neq \varnothing\big) \quad \Longleftrightarrow \quad \big(\forall S \in \mathcal{F}_\ell^{\mathrm{NF}} :\ P \cap S \neq \varnothing\big).$$

The forward direction uses subfamily inclusion. For the reverse direction, a successful edit avoiding P would normalize to one that still avoids P. Both families include all permitted costs, not just minimum-cost witnesses. If the initial report is already the target, the empty edit set occurs in both families and makes blocking impossible. The argument uses a singleton target; maintaining both labels requires a separate condition.

B.4 The cost boundary of survivor alignment

This section proves Proposition 1: first the weighted extension, then the sharp boundary when only missingness is reweighted. All statements concern resolution from a feasible ambiguous center. Let $\mathcal{H} \subseteq \{0,1\}^r$ have a fixed binary label, let $o_0 \in \{0,1,\mathtt{M}\}^r$, and let $E \subseteq [r]$ be editable. A *survivor-aligned record* has some compatible world v such that every newly written nonmissing value agrees with v. Equivalently, it aligns a set R of initially missing or mismatching editable slots to v, and deletes a disjoint set D of initially recorded editable slots. Values kept unchanged may disagree with v. The survivor certifies final support feasibility; factual containment follows under the cumulative validity condition.

For positive slot weights c_j and edit weights κ_j, write

$$d_c(w,o) = \sum_{j=1}^{r} c_j \mathbf{1}[o_j = \mathtt{M} \text{ or } o_j \neq w_j], \qquad \delta_\kappa(o_0,o) = \sum_{j=1}^{r} \kappa_j \mathbf{1}[o_{0j} \neq o_j].$$

Set $\Omega_b^c(o) = \{w \in \mathcal{H} : d_c(w,o) \leq b\}$. The same κ_j applies to either replacement symbol. Budgets a, b may be nonnegative reals; there are finitely many records, so the relevant minima exist whenever the feasible sets are nonempty. Rational weights give a finite encoded optimization problem. Write $c(S) = \sum_{j \in S} c_j$ and analogously $\kappa(S)$. Define

$$J_w = \{j \in E : o_{0j} = w_j \in \{0,1\}\}, \quad N_v = E \setminus J_v, \quad K = \{j \in E : o_{0j} \neq \mathtt{M}\}.$$

For $R \subseteq N_v$, let $o^{v,R}$ denote the record aligning R to v.

Proposition 4 (Within-slot equality, with heterogeneous weights)**.** *For a feasible ambiguous center, the weighted report-change radius is*

$$\min_{v \in \mathcal{H},\ R \subseteq N_v,\ D \subseteq K \setminus R} \kappa(R) + \kappa(D)$$

subject to

$$\begin{aligned} \kappa(R) + \kappa(D) &\leq a, \\ d_c(v,o_0) - c(R) + c(D \cap J_v) &\leq b, \\ c(D \cap J_w) &> b - d_c(w, o^{v,R}) \quad \textit{for every } w \in \Omega_b^c(o^{v,R}) \textit{ with } y(w) \neq y(v). \end{aligned}$$

An empty feasible set means no nonempty report change within the specified budget. The result holds for arbitrary editable domains and does not require a persistent core or an initially compatible survivor.

Proof. Take any successful record o and any $v \in \Omega_b^c(o)$. At an initially missing slot filled with $1 - v_j$, restore the missing symbol. At an initially recorded slot $o_{0j} = v_j$ flipped to $1 - v_j$, replace that final value by missing. In either replacement the distance contribution to v stays c_j. For every world with bit v_j it also stays c_j; for a world with bit $1 - v_j$ it increases from zero to c_j. The first replacement reduces edit cost from κ_j to zero, and the second preserves κ_j. Therefore the transformed record retains v, excludes all formerly excluded worlds, and costs no more. Applying these coordinatewise operations produces precisely the stated alignment/deletion form.

Alignment lowers v's distance by $c(R)$. Since $D \cap R = \varnothing$, the aligned record still equals o_0 on D, and for every world w,

$$d_c(w, o^{v,R,D}) = d_c(w, o^{v,R}) + c(D \cap J_w).$$

Deletion cannot admit a previously incompatible world. Consequently the retention and strict exclusion inequalities are necessary and sufficient for the proposed final support to be nonempty and have label $y(v)$. The edit cost is $\kappa(R) + \kappa(D)$. Every unrestricted solution thus has a restricted solution of no greater cost, while every feasible restricted solution is itself an unrestricted solution. The optima coincide. Initial ambiguity rules out a zero-cost feasible solution. □

The strict inequality is essential for general weights: one may replace it by an integer "$+1$" form only after establishing the appropriate integer lattice. Weights do not automatically preserve the unweighted branch bound. Let $\eta = \max_{j \in E} c_j/\kappa_j$ and $c_{\min} = \min_j c_j$, for nonempty E. A survivor obeys

$$d_c(v, o_0) \le b + a\eta, \qquad |N_v| \le (b + a\eta)/c_{\min}.$$

The first inequality follows because changing a set of slots decreases distance by at most its total c weight; the second follows by charging at least $c_{\min}$ for each initial nonmatch. Thus bounds on weight ratios can control alignment enumeration, but $|\mathcal{H}|2^{a+b}$ does not follow for unrestricted weights. For instance, $c_j = \kappa_j = 1/r$ at an all-missing center with $a = b = 1$ permits 2^r alignment subsets per survivor. The integer distance-profile bound in Theorem 4 uses the unit-distance lattice; real-weight profiles require a separate state count.

Proposition 5 (Exact universal boundary under unequal missing cost)**.** *Fix $\lambda \ge 0$ and use*

$$d_\lambda(w, o) = \lambda\,|\{j : o_j = \mathtt{M}\}| + |\{j : o_j \in \{0,1\},\ o_j \ne w_j\}|,$$

with unit edit costs. Over all finite Boolean labeled domains, feasible ambiguous centers, editable sets and budgets, restricting to survivor-aligned records preserves the exact radius and bounded solvability if and only if $\lambda = 1$. For each $\lambda \ne 1$, failure occurs even at a complete center with all coordinates editable and matched positive integer budgets.

Proof. For $\lambda = 1$, Proposition 4 applies with $c_j = \kappa_j = 1$.

For $\lambda > 1$, take $\mathcal{H} = \{00, 01\}$ with respective labels zero and one, $o_0 = 00$, $E = [2]$, and $a = b = 1$. The initial distances are $(0, 1)$, so the initial report is ambiguous. Record 10 has distances $(1, 2)$ and is definite zero, proving radius one. Its sole survivor 00 disagrees with the newly written first bit, so it is not survivor-aligned. The other one-edit records are 01, $\mathtt{M}0$ and $0\mathtt{M}$. The first is ambiguous, with distances $(1, 0)$; the latter two are infeasible, with distances $(\lambda, \lambda + 1)$ and (λ, λ) respectively. No restricted solution exists within budget.

For $0 \le \lambda < 1$, let $t = \lceil 1/(1 - \lambda) \rceil$, so $t \ge 1$ and $t\lambda \le t - 1$. Take

$$r = t + 1, \quad \mathcal{H} = \{x = 0^{t+1}, z = 0^t 1\}, \quad y(x) = 0, \quad y(z) = 1,$$
$$o_0 = x, \quad E = [r], \quad a = b = t.$$

Both worlds are initially compatible. In any record, let p be the common distance contribution from its first t slots. If the last slot is missing, the two distances agree and the support cannot be nonempty monochromatic. If that slot is recorded, the distances are $(p, p + 1)$ or $(p + 1, p)$; a nonempty monochromatic support therefore requires $t - 1 < p \le t$. Each edited common slot contributes at most one because $\lambda < 1$. Fewer than t edits cannot meet this condition. Flipping all first t zeros to one gives distances $(t, t + 1)$, so the unrestricted radius is exactly t.

Both possible survivors have zero in every common slot. A survivor-aligned record can only retain or delete those slots, hence its common contribution is at most $t\lambda \le t - 1$. If its final slot is recorded, both worlds remain compatible; if that slot is missing, their distances are equal. It never has nonempty monochromatic support. In particular, no restricted solution exists within the matched budget. This covers $\lambda = 0$ as well. □

Both two-world counterexamples admit the causal interface used in the paper: for worlds x, z labeled zero and one, take $A = 1$ and $h = (A \wedge [u = z]) \vee [u = x]$. Disabling the sole candidate A prevents harm exactly at z; no extra uncertain evidence slot is required.

Unequal charges change the distances within this edit restriction. If R_M and R_X are its initially missing and initially recorded mismatching alignments, then

$$d_\lambda(v, o^{v,R,D}) = d_\lambda(v, o_0) - \lambda|R_M| - |R_X| + (\lambda - 1)|D| + |D \cap J_v|,$$
$$d_\lambda(w, o^{v,R,D}) = d_\lambda(w, o^{v,R}) + (\lambda - 1)|D| + |D \cap J_w|.$$

For $\lambda < 1$, deleting a retained value that mismatches w decreases its distance; opposite worlds initially outside $\Omega_b^\lambda(o^{v,R})$ may enter. Thus the original compatibility filtering is also unjustified. For $\lambda \geq 1$, deletion remains distance-monotone, but monotonicity alone does not guarantee preservation of a survivor when $\lambda > 1$.

B.5 Cumulative validity and finite-budget invariance

Proof of Corollary 1. Equation (6) puts $u^\star$ in the persistent core. Every successful resolution therefore has label $y(u^\star)$; a core world of the other label would make resolution impossible. Given any successful record, undo fills of initially missing slots and replace recorded-bit flips by missing. Both transformations weakly increase every world's distance, so no opposite-label world reenters. The factual world remains and edit cost does not increase. Applied to an optimum, the deletion-only result has equal cost. Indeed, an optimal record cannot contain a fill, whose removal would strictly save an edit; it can contain a recorded-bit flip with an equal-cost deletion alternative. Several edit types can therefore attain the optimum. □

For example, with $\mathcal{H} = \{000 : 0, 001 : 1\}$, $o_0 = u^\star = 000$, $b_0 = 0$ and $a = b = 2$, both 1M0 and MM0 attain radius two. The first witness is mixed, but mixing is not necessary. With incomplete center M000, worlds $\{0000 : 0, 0001 : 1\}$, $b_0 = 1$, $a = 1$ and $b = 2$, erasing a common recorded zero resolves ambiguity. Cumulative validity thus allows resolution, while constraining its label and optimal edit form.

Proposition 6 (Finite-budget cost invariance)**.** *For a complete o_0, nonnegative integers a, b and $0 \leq \lambda < 1$, if $a(1 - \lambda) < 1$, then every $o \in \mathcal{B}_a(o_0)$ satisfies $\Omega_b^\lambda(o) = \Omega_b^1(o)$, where $\Omega_b^\lambda(o) = \{w \in \mathcal{H} : d_\lambda(w, o) \leq b\}$. All feasible query radii and initially ambiguous normal-form optima are unchanged.*

Proof of Proposition 6. Let $\varepsilon = 1 - \lambda$ and $m(o)$ count missing slots. Completeness of o_0 gives $m(o) \leq a$. Hence

$$d_\lambda(w, o) = d_1(w, o) - \varepsilon m(o), \qquad 0 \leq \varepsilon m(o) < 1.$$

A unit-distance-compatible world remains compatible. Otherwise integer b and d_1 give $d_1(w, o) \geq b+1$, and consequently $d_\lambda(w, o) > b$. Supports therefore agree everywhere in the edit ball. Feasible reports, attaining records, radii and bounded unflippability agree as well. Whether an edit is survivor-aligned depends on its symbols and a compatible agreeing survivor; these also agree. Completeness of the unit-cost normal form at an initially ambiguous center transfers to λ. □

A budget-dependent sharp boundary. For fixed $0 \leq \lambda < 1$, let $t = \lceil 1/(1 - \lambda) \rceil$. The guarantee covers integer budgets $a < t$. At the first excluded budget $a = b = t$, use x, z from Proposition 5. Deleting their first t common slots gives distances $(t, t+1)$ at unit cost but $(t\lambda, t\lambda+1)$ at λ, admitting both worlds in the latter case. The same proposition proves failure of normalization. Thus both guarantees are sharp uniformly over domains. The unrestricted radii in this family are nevertheless both t; different supports need not change the numerical radius.

For an incomplete center with m_0 missing entries, the same reasoning gives support invariance whenever

$$(1 - \lambda)\big(m_0 + \min\{a, |E \cap \{j : o_{0j} \neq \mathtt{M}\}|\}\big) < 1.$$

The parenthesis is the largest attainable missing count. In the complete case, a may accordingly be replaced by $\min(a, |E|)$.

B.6 Ordinary certificates as analytic references

A coordinate permutation preserves all target radii when applied consistently to worlds, records, labels, permissions and costs. Independently editable copies or subfields define a different corruption process. A grouped encoding preserves the original query when its legal records, edits and charges correspond bijectively and preserve both distances and labels. Suppose $\mathcal{H} = \{0,1\}^r$, all slots are editable, $o_0 = \mathtt{M}^m 0^{r-m}$ and $b = m$. Let the nonconstant label f depend only on the first $q \le m$ roots. A forcing certificate is a partial assignment of these roots on which every completion has the same label. Let $t \ge 1$ be the smallest certificate size for either label. Then

$$\rho^* = 2t \quad \text{if } a \ge 2t \text{ and } r - m \ge t; \qquad \text{otherwise } \rho^* = \infty. \tag{17}$$

Proof. Initially known roots are label-irrelevant. Flipping their recorded values can be undone without changing projected support on relevant roots, since nuisance values can match either record. Describe the remaining edits by k filled relevant slots, u filled nuisance slots and p deleted known nuisance slots. The missing count is $m - k - u + p$, leaving retained-error allowance $e = k + u - p$. Nonempty support requires $e \ge 0$. Projected support consists of all relevant assignments differing from the k fixed values in at most e positions. If $e \ge k$, every relevant assignment appears, contradicting a definite label. Thus $e < k$.

From a definite record, retain any $k - e$ fixed relevant values and free the other e. Every completion differs in at most e of the originally fixed positions and has the same label, so it is a forcing certificate. Hence $t \le k - e = p - u$. In particular $p \ge t$ requires $r - m \ge t$, and

$$k + u + p = 2(k + u) - e \ge 2t + e + 2u \ge 2t.$$

Conversely, fill a minimum t-literal certificate and delete any t initially known slots. The missing count stays $m = b$, so no retained-value errors are allowed. Support is nonempty and every world obeys the certificate, at cost $2t$. This proves both the radius and the impossibility conditions. □

For a projection or an AND/OR function, the minimum certificate for either label has size one. For parity on q roots it has size q; a three-input multiplexer has minimum size two. Increasing the number of nuisance roots enlarges the world cube without making the label function intrinsically harder. These are analytic examples of the certificate relation. The full-cube projection example in Figure 1 is also causally realizable: take an endogenous action $A = 1$ and $h = A \vee \neg U_1$. Every root assignment is harmful; disabling A prevents harm exactly when $U_1 = 1$.

C Outer Complexity Proofs

C.1 Radius-only NP-hardness

Theorem 5 (Radius-only NP-hardness)**.** *Deciding $\rho^*(o_0; B, B) \le k$ is NP-hard with one candidate, no freezes, $\tau = 0$, every evidence slot editable, a polynomial-size explicit harmful-world table and constant-time local causal scores. For a nonempty simple graph on $n \ge 3$ vertices, a construction with $r = n + 1$ and $B = n - 1$ has radius equal to its minimum vertex-cover size.*

Proof of Theorem 5. Construction. Roots are one evidence slot x_v per vertex plus one further slot g, all corruptible, so $r = n+1$. Equations, at fan-in two: $P = \bigwedge_{v \in V} x_v$; $Q = \bigvee_{e \in E} \big[x_V = w_e^V \wedge g = 0\big]$, where w_e^V assigns 1 exactly to the endpoints of e; and $h = (g \wedge P) \vee Q$. Register $C = \{g\}$, $a_g = 0$, $F = \varnothing$, $\tau = 0$, $B = n - 1$, and the clean observation $o_0 = u_0 = (1^n, 1)$.

Worlds and scores. The two disjuncts of h are disjoint and force their contexts, so the harmful set is exactly $\{u_0\} \cup \{w_e : e \in E\}$ with $w_e = (w_e^V, 0)$. At u_0 the candidate is active and setting g to its safe contrast gives $h = 0$, so $\{g\}$ is a minimal cause and $s_g(u_0) = 1$; at w_e the candidate is inactive ($z_g^* = 0 = a_g$), so $s_g(w_e) = 0$ is an exact zero by Equation (13), with no replay required. Thus on every explicitly listed harmful world the unified scorer is simply $s_g(u) = u_g$, a constant-time coordinate lookup. Moreover, each w_e agrees with u_0 exactly on the two endpoint slots of e, hence $d(w_e, o_0) = n - 1 = B$: the clean compatible set is all of $\mathcal{H}$ and, because $E \neq \varnothing$, the clean report is ambiguous.

Edit calculus. For a complete clean observation the explanation distance is additive across slots. On a slot where a world agrees with the clean bit, deleting or replacing that slot each add exactly one; on a disagreeing slot, deletion adds zero and replacement subtracts one. So deletion pointwise dominates replacement; $d(u_0, o) = |S|$ for every edit set S of either kind; and for an all-deletion S, $d(w_e, o) = (n-1) + |S \cap e|$.

Correctness. For every $|S| \le B = n-1$, the world u_0 stays compatible, so the attacked report is never infeasible and never definite-not-above, and a flip occurs exactly when every w_e is evicted. With deletions, w_e is evicted iff S meets e, so all are evicted iff S is a vertex cover; by dominance, if any mixed-kind S evicts every w_e, its all-deletion variant of equal cost does too. Hence $\rho^*(o_0; n-1, n-1) = \tau_{\mathrm{vc}}(G)$. For a nonempty simple graph, $1 \le \tau_{\mathrm{vc}}(G) \le n-1$, so this radius lies in the registered positive domain. The decision equivalence $\rho^*(o_0; n-1, n-1) \le k$ iff $\tau_{\mathrm{vc}}(G) \le k$ follows for $0 \le k \le n-2$. Vertex Cover remains NP-hard on this promise. Starting from its standard simple-graph form, preprocess empty or $n < 3$ graphs directly, treat a negative bound as no, and observe that $k \ge n-1$ is yes because $V \setminus \{v\}$ covers every edge. Map each resolved case to either the fixed promised yes-instance consisting of one edge and one isolated vertex with $k = 1$, or the fixed promised no-instance consisting of a triangle with $k = 1$. Every unresolved instance is nonempty, has $n \ge 3$, and satisfies $0 \le k \le n-2$, so applying the construction completes the reduction. □

C.2 Budget-parameterized hardness

Proof of Theorem 3. Source promise. Parameterized Independent Set is W[1]-complete (Downey & Fellows, 1995, Theorem 3.1). We may require a simple graph $G = (V, E_G)$ with no isolated vertices and $2 \le k < |V|$. Given an ordinary instance with $2 \le k \le n$, add a universal vertex. The new graph has no isolated vertices and $k < n+1$; every independent set of size at least two excludes the new vertex, preserving the answer and k. Trivial original cases may be decided directly and mapped to $(P_3, 2)$ or $(K_3, 2)$ as fixed yes/no targets. This preprocessing is parameter preserving up to the computable bound $\max(k, 2)$.

Explicit audit. Write $n = |V|$, and create n uncertain, editable coordinates, complete factual center $o_0 = 0^n$, and

$$\mathcal{H} = \{0^n\} \cup \{\mathbf{1}_e : e \in E_G\}, \qquad y(0^n) = 1, \quad y(\mathbf{1}_e) = 0, \qquad B = k.$$

The table has $1 + |E_G|$ rows and $O(n(1 + |E_G|))$ bits. Proposition 7 proves that its charged-radius query is feasible exactly when G has an independent set of size at least k, and that every bounded finite answer equals k. That proof supplies the deletion normalization and the lower-cost impossibility cases.

Causal realization without extra evidence slots. Take uncertain roots x_v, a constant endogenous action $A = 1$, and Boolean equations

$$Z(x) = \bigwedge_v \neg x_v, \qquad Q(x) = \bigvee_{e \in E_G} [x = \mathbf{1}_e], \qquad h = (A \wedge Z(x)) \vee Q(x).$$

Register the sole candidate action A, safe value zero, empty freeze domain, and strict threshold $\tau = 1/2$. Factual harmful worlds are exactly the specified table. At 0^n, disabling A prevents harm, so its singleton is a minimal cause with score one. At each edge world, $Q = 1$ regardless of A, so its score is zero. No empty intervention prevents factual harm. To realize the root-candidate interface, use the trusted selector σ_A, factually zero and safely one, and set $A = \neg\sigma_A$ as in Proposition 3. The selector is outside the uncertain evidence projection and attack universe. Thus there are still exactly $r = n$ evidence slots and $B = k$; adding an editable selector would be a different construction. Binary fan-in trees implement all conjunctions, disjunctions and edge equalities with $O(n(1+|E_G|))$ gates. The resulting SCM is finite, total and acyclic; a valid one-bag decomposition may be supplied without a width bound.

The transformation is polynomial, and each target parameter B, k and $B + k$ is bounded solely by the source parameter k. It establishes the stated W[1]-hardness for each choice of parameter. Stored explicit labels have constant-time access. In the causal realization, testing $x = 0^n$ on a harmful root or replaying the circuit takes polynomial time in this realization. □

C.3 THE SAME TABLE SUPPORTS DIFFERENT PARAMETERIZED QUERIES

For a finite labeled domain $(\mathcal{H}, y)$ and distinguished world x, an *ordinary anchor* is a coordinate set A such that every $w \in \mathcal{H}$ agreeing with x on A has label $y(x)$. This is the coverage-zero instance of MINIMUM ANCHOR (Ordyniak et al., 2023, Section 2); coverage counts other same-label examples and is not required here.

Proposition 7 (Same-table anchor/radius contrast)**.** *Let $G = ([r], E_G)$ be a simple graph without isolated vertices, and let $2 \le k < r$. Set*

$$\mathcal{H}_G = \{0^r\} \cup \{\mathbf{1}_e : e \in E_G\}, \qquad y(0^r) = 1, \quad y(\mathbf{1}_e) = 0.$$

On this explicit labeled table, an ordinary anchor of size at most k for 0^r, with coverage requirement zero, exists exactly when G has a vertex cover of size at most k. Under unit charged distance, full editability, center 0^r and $a = b = k$, a report-changing record exists exactly when G has an independent set of size at least k. In that case the radius is k; otherwise it is unflippable within budget k.

Proof. An edge world disagrees with 0^r exactly at its endpoints. Hence A is an anchor precisely when $A \cap e \neq \varnothing$ for every edge e, the vertex cover condition. The only same-label row is the distinguished row itself, so its coverage under the cited definition is zero.

For the radius query, all rows are initially compatible because $k \ge 2$. Every allowed edited record remains within distance k of 0^r: each changed zero, whether replaced by one or by missing, contributes exactly one. Thus a report change must exclude all edge worlds while retaining the zero world. Replacing every edited one by missing preserves cost and distance to 0^r, and weakly increases distance to every edge world. It therefore suffices to consider deletion sets D, with $t = |D| \le k$. Their edge distance is

$$d(\mathbf{1}_e, o^D) = 2 + t - |D \cap e|.$$

If $t \le k - 2$, every edge world remains compatible. If $t = k - 1$, excluding every edge would require $D \cap e = \varnothing$ for every edge. But D is nonempty because $k \ge 2$, and no vertex is isolated. If $t = k$, every edge is excluded exactly when $|D \cap e| \le 1$ for all e, which says that D is independent. An independent set of size at least k contains one of size exactly k. These observations prove both equivalences and the exact radius. □

The first decision problem admits the standard 2^k branching algorithm: choose an uncovered edge and branch on selecting either endpoint. This is also the distance-two special case of the FPT result parameterized by anchor size and maximum opposite-example difference (Ordyniak et al., 2023, Theorem 7). The second problem is W[1]-hard in k by INDEPENDENT SET (Downey & Fellows, 1995). The source promise does not remove that hardness: after handling trivial cases, adding a universal vertex removes all isolated vertices, enforces $k < r$, and preserves independent sets of size at least two and the parameter k. The causal realization in Appendix C uses the same uncertain evidence slots and labels; no new selector enters the edit universe.

At a successful size-k deletion, the retained complement is an ordinary anchor of size $r - k$; this complementation does not preserve the parameter k. The comparison holds the data and small parameter fixed while changing the query contract. The restriction that all opposite rows lie at distance two suffices for anchor FPT in k, but not for charged-radius FPT in k unless $\mathrm{FPT} = \mathrm{W}[1]$. The distinction concerns the shared-budget exclusion constraint on this table.

D EXACT RADIUS ON A JOINT LAYERED PROGRAM

Let $Q_0, \dots, Q_r$ be the layers, q_0 the initial node, $\eta_j : Q_{j-1} \times \{0,1\} \to Q_j$ the total transition functions, and $\ell_{\mathrm{out}} : Q_r \to \{\bot, 0, 1\}$ the terminal labels. Every path reads all coordinates in the same order. Here $\eta(u_{1:j})$ denotes the node reached from q_0 after that binary prefix. Put $A = \min(a, |E|)$, $B = \min(b, r)$, $\nu = \min\{B + 2, 2A + 1\}$ and $c(x, s) = 1$ for $s = \mathtt{M}$, otherwise $c(x, s) = \mathbf{1}[x \neq s]$. For an observed prefix, its profile is

$$p_j(q) = \min\left\{ B + 1, \min_{u_{1:j}:\, \eta(u_{1:j}) = q} \sum_{i=1}^{j} c(u_i, o_i) \right\}. \tag{18}$$

An empty inner minimum is infinity. Initialize $p_0(q_0) = 0$ and all other entries to $B + 1$. For a next observed symbol s, the exact update is

$$(T_{j,s}p)(q') = \min\left\{B+1, \min_{q,x:\eta_j(q,x)=q'}\{p(q) + c(x,s)\}\right\}. \tag{19}$$

The available symbols are $\{o_{0j}\}$ at protected coordinates and $\{\mathtt{M}, 0, 1\}$ otherwise. Store the least observation-edit cost $C_j(p)$ attaining each profile. Starting with $C_0(p_0) = 0$, update

$$C_j(p') = \min_{p,s:\, T_{j,s}p=p'}\{C_{j-1}(p) + \mathbf{1}[s \neq o_{0j}]\}, \tag{20}$$

discarding costs above a. The final supported labels are $\Lambda(p) = \{\ell_{\text{out}}(q) \in \{0,1\} : p(q) \leq B\}$. First evaluate the initial record to obtain Λ_0; if it is empty, report infeasibility. Otherwise minimize $C_r(p)$ over $\varnothing \neq \Lambda(p) \neq \Lambda_0$. No such state means bounded unflippability. Stored predecessors recover an attaining record.

Proof of Theorem 4. Induction on the input layer establishes (18): every binary prefix has a unique last transition, and (19) minimizes over all such predecessors. Truncation cannot remove a subsequently feasible prefix, because charged costs are nonnegative. Two observed prefixes yielding the same profile have identical profile extensions for every common suffix. The more expensive prefix is therefore dominated. A second induction proves that (20) retains exactly the minimum edit cost attaining each profile. At the final layer, a state has an in-budget world exactly when its profile entry is at most B; thus $\Lambda(p)$ is the exact harmful-label support. The final test enforces both feasibility and a changed report, proving optimality.

Let $\bar{p}_j$ be the profile of the center prefix $o_{0,1:j}$ and put $A_j = \min(a, |E \cap [j]|)$. Any allowed prefix differs from that center in at most A_j slots. For a fixed binary path, changing one observed symbol changes its charged distance by at most one. The paths reaching a node are independent of the observation, so taking minima over that same nonempty path set preserves the bound A_j. Truncation at $B + 1$ also preserves it; unreachable nodes have entry $B + 1$ in both profiles. Hence

$$\|p_j - \bar{p}_j\|_\infty \leq A_j.$$

Each entry lies in $[\bar{p}_j(q) - A_j, \bar{p}_j(q) + A_j] \cap \{0, \ldots, B+1\}$, giving at most $\min\{B+2, 2A_j+1\}^{|Q_j|} \leq \nu^w$ profiles at layer j. For each profile and at most three observed symbols, a transition scans $2w$ binary edges. The time bound is therefore $O(rw\nu^w)$ arithmetic operations. Keeping only the least edit cost per profile avoids a further factor in a. Two-layer working space is $O(w\nu^w)$; retaining all predecessors for direct witness recovery uses $O(rw\nu^w)$ space. Integers require $O(\log(r+2))$ bits after capping the attack budget at $|E|$. The same recurrence is thus FPT in (w, a) and (w, b) at arbitrary centers and editing permissions. □

The width is that of a complete joint program for harmful-domain membership and threshold labels. Completing skipped levels or finding a suitable variable order can change this width. Same-order complete programs for the two predicates admit a product construction of width at most $w_H w_y$; an arbitrary succinct SCM need not admit a small product. Compilation is outside the theorem's cost. The recurrence is a specialization of standard min-plus and bounded-width methods to this charged-distance query.

Product construction for a fixed plan. Let the factual and counterfactual transition functions at root layer j be T_j^F and T_j^C. On input u_j, update the pair by $(q_F, q_C) \mapsto (T_j^F(q_F, u_j), T_j^C(q_C, u_j))$. The state spaces include all states reachable under the registered intervention as well as the factual execution. Both programs must read every root in the same order, including initial-state parameters. On the final pair output $\perp$ if factual harm is zero, and otherwise $1 - h_C(q_C)$. With exact transition and harm tables this is precisely the strict-half singleton causal label when no freezes are allowed. Any extra admissibility constraint must be included in the state, or its monitor width multiplies the bound. If the factual/intervened machines have maximum layer sizes $s_F, s_C \leq s$, the product has width at most $s_F s_C \leq s^2$ and is constructed in $O(rs^2)$ time. Applying the profile recurrence proves the stated radius bound from these exact state machines.

Length illustration. The frozen joint-width grid uses the no-adjacent-ones domain and labels $\sum_j u_j \equiv 1 \pmod p$, for $p = 2, 3$. Each root is label-relevant by comparing the all-zero word with the word having only that root set. Joint complete widths are five and seven. At lengths 8, 32, 128 and 512, three profiles cover partially missing centers and protected positions. All twelve radii are two, checked by an independent attainable record and one-edit lower bound; the eight-root cases are also checked by full observation search. At length 512 the recorded inference times are 0.016, 0.068 and 0.014 seconds, with at most 29, 119 and 80 profiles per layer respectively. These radius-two examples measure inference on supplied narrow programs, illustrating the representation bound through length 512.

E Relationship to Existing Explanation Queries

Let $y : \mathcal{H} \to \{0, 1\}$ be the fixed world label and $G_b(o) = \{y(w) : w \in \Omega_b(o)\}$. The radius is a constrained contrastive query of the set-valued map G_b, with infeasible records excluded. Its input is a record, and its output aggregates labels over the charged support.

Retaining compatible labels follows partial identification under uncertain or corrupted observations (Horowitz & Manski, 1995; Manski, 2003). Consistent query answering retains conclusions shared across database repairs (Arenas et al., 1999); epistemic causal explanation reasons across possible contexts (Halpern & Pearl, 2005b). These connections concern aggregation over alternatives, not an identity of the optimization problems. Database repairs are selected by consistency and a minimal-change criterion. Here $\Omega_b(o)$ contains every world within the charged bound, whether or not it is a closest explanation of the record. The radius then changes o and recomputes this entire support. A data-complexity correspondence would therefore have to preserve the permitted repairs, their minimality criterion, the outer edit objective, and the representation of the world domain.

Database causality studies which tuple changes cause a query answer and how responsibility depends on contingencies (Meliou et al., 2010). In our causal instance, intervention inference supplies $y(w)$ separately in each world. Record-edit optimization keeps that map fixed and changes which worlds remain supported. The hardness results isolate this outer task even when computing the local label is easy.

Wildcard partial inputs. Béjar et al. (2023, Equation 9) map partial feature assignments to possible class sets and define explanations for such sets. Wildcard completion fixes recorded values and leaves unspecified values free. Here a record with m missing slots permits up to $b - m$ errors among its retained values. Deleting a recorded slot both frees its value and reduces that residual allowance. This coupling can contract support. Filling a slot reverses the allowance change; the normal form characterizes how to combine such edits with a surviving world. Input constraints and symbolic completion methods from the prior framework remain applicable tools.

Abductive/contrastive explanations and covering. For a complete factual center and deletion set D with $|D| \le b$, write $R = [r] \setminus D$. An opposite world with discrepancy set Δ_w has distance $|D| + |R \cap \Delta_w|$. Retaining the factual label is therefore equivalent to

$$|R \cap \Delta_w| > b - |D| \quad \text{for every opposite } w. \tag{21}$$

Equivalently, the retained factual features still form a sufficient explanation after any further removal of up to $b - |D|$ retained positions. This is a robust abductive condition with a residual tolerance coupled to deletion size. Each multicover inequality is also a compact representation of ordinary hitting-set constraints: $|D \cap J| \ge q$ means that D meets every $(|J| - q + 1)$-subset of J. Covering and oracle-guided optimization supply established solution methods (Ignatiev et al., 2021; La Malfa et al., 2021).

Small-parameter hardness on explicit data. The same-table comparison in Proposition 7 instantiates the distance-two tractable anchor class of Ordyniak et al. (2023, Theorem 7), while retaining W[1]-hard charged resolution at the same parameter. It distinguishes retained trusted features from charged deletions. The survivor in our normal form is a world, not an anchor feature set.

From world programs to record reports. Ordyniak et al. (2024, Corollary 12) obtain fixed-parameter tractability for standard explanations on complete common-order majority-vote OBDD

ensembles, parameterized by ensemble size and member width. Their Theorem 14 also gives a polynomial algorithm for a minimum local contrastive explanation on one supplied OBDD. Those inputs already classify feature vectors. Our supplied joint program instead classifies binary worlds as inadmissible, zero or one. Computing G_b additionally aggregates all worlds within charged distance b of a ternary record. At its terminal layer, a distance profile determines all four possible reports, including empty support. The outer search enforces E, edit budget a and the nonempty changed-report condition.

Thus a representation of G_b would turn the radius into a constrained contrastive query on records. Theorem 4 bounds the intervening aggregation and search directly: a world-program layer of width w needs at most $\min\{B+2, 2A+1\}^w$ profiles over the allowed edit ball. The $B+2$ count comes from saturated distances; the $2A+1$ window comes from the center and edit budget. This explicit bound starts from the joint world representation. Applying a standard explanation algorithm after encoding ternary records would also require constructing its record classifier and tracking the resulting width, edit permissions and feasibility constraint.

F Symbolic Search and Independent Certificates

The implementation instantiates labels by a single registered safe-action intervention with no freezes. It encodes a factual and an intervened copy of the input acyclic Boolean SCM; harm is evaluated on the factual copy and singleton prevention on the intervened copy. Thus the tested label is obtained from the factual/intervened model semantics. Richer local labels require their corresponding exact oracle and checker. The circuit extends y arbitrarily outside $\mathcal{H}$; domain membership excludes these values from every report, so this total extension leaves all queries unchanged.

F.1 Master and counterexample oracle

For each positive edit bound $k \le a$ and target label ℓ, a master jointly chooses a record o and harmful survivor v satisfying $y(v) = \ell$, $d(v,o) \le b$, $\delta(o_0, o) \le k$, and the protected-coordinate constraints. The unrestricted route uses arbitrary ternary records. The normal-form route additionally requires every newly recorded bit to agree with v; deletion of initially recorded positions is allowed. Keeping an old mismatching bit is also allowed. These are exactly the coordinate possibilities in Theorem 2, not an assumption that all observed values equal the survivor.

For an initially ambiguous support, a proposed record is checked by the original-SCM oracle for a compatible opposite-label world. Each returned world w yields the necessary master cut $d(w,o) > b$. The master may change its survivor later; the cut remains necessary because the target label stays fixed. The oracle searches the whole harmful domain, including worlds outside the original support. Oracle unsatisfiability proves that the nonempty support is definite; master unsatisfiability excludes that target at the current bound. Both target labels must be excluded at every lower bound before a finite radius is reported. If neither succeeds through a, the answer is bounded unflippability.

For an initially definite support, admitting a survivor of the other label already changes the report. The original label need not disappear, and the implementation does not require an opposite-world exclusion in this branch. An initially empty support is reported as infeasible before radius search.

F.2 Shared certified cube generalization

A Boolean cube C fixes coordinates F_C and leaves the rest free. The same generalization procedure is used in both routes. Starting from a checked opposite world, each removal of a cube literal is admitted only after an original-SCM unsatisfiability proof establishes that every completion of the enlarged cube is harmful and has the opposite label. The counterexample query is $C(w) \wedge (h(w) = 0 \vee y(w) \neq 1 - \ell)$, where ℓ is the master target. Checking the label only conditionally on harm would be insufficient.

For every record, the nearest cube world's charged distance is

$$\min_{w \in C} d(w,o) = |\{j : o_j = \mathtt{M}\}| + |\{j \in F_C : o_j \neq \mathtt{M},\ o_j \neq C_j\}|. \tag{22}$$

Every missing slot costs one; recorded free coordinates can be matched at zero cost, and fixed mismatches are unavoidable. This proves the formula. Requiring its value to exceed b excludes the

whole certified opposite cube and is a necessary cut for a target-definite support. The final opposite-world oracle is still required. The independently reconstructed proof of cube homogeneity and the distance formula make these cuts valid in lower-bound masters. Both routes share the generalization procedure, literal order and proof requirements.

F.3 Soundness, completeness and finite termination

Every master proposal retains a target-label world, ensuring nonempty support. Every cut follows from an actual opposite world, so no successful record of that target is removed. Unsatisfiable masters therefore prove lower bounds for the full problem despite containing only a subset of the necessary cuts. A feasible proposal with an unsatisfiable opposite oracle has the claimed label. This proves soundness of accepted witnesses and lower bounds.

For unrestricted search, any successful record and one of its surviving worlds satisfy the master and every necessary cut. For normal-form search, Theorem 2 supplies such a pair of no greater cost. For an initially definite support, the aligning construction in Theorem 1 supplies an optimal opposite-label survivor in the restricted domain as well. Thus an optimal success cannot be excluded by either master. A failed oracle proposal adds a cut violated by that record, preventing its repetition at the same target. There are finitely many records and worlds. With complete solver calls and no resource interruption, the process terminates and establishes a finite optimum or bounded unflippability; the number of records may be exponential in the input dimension.

F.4 Independent proof obligations

The independent certificate checker reconstructs the SCM, record/survivor relation, cardinalities and normal-form constraints without importing the producer or SAT solver. It checks satisfying assignments against both the reconstructed CNF and direct SCM evaluation, and unsatisfiability proofs with an independent RUP checker. A submitted CNF hash or solver status is insufficient. The transcript must link every oracle witness to its cut and every final answer to all required lower-budget and target-label obligations. Omitted obligations, changed labels or illegal edits invalidate the certificate. Resource-limited prefixes remain unsolved even if their existing proof steps are valid.

The two routes use the same label semantics, resource limits and output requirements. Timings distinguish producer and checker workers; embedded encoding, solver and proof-writing timers are nested and cannot be added as if they were disjoint. Native proof production is included in solving. All attempts, including failures, are retained in the artifact. Implementation and admission details for the supplied Normal route are documented in the code supplement.

G Computational Study

G.1 General-center inputs and measurement protocol

The grid crosses two generators, $r \in \{8, 12, 16\}$ and three profiles. For family $j = 0$ (dependency) or 1 (layered) and profile $q \in \{0, 1, 2\}$, the seed is $420900 + 1000j + 10r + q$. Its shuffled order determines paired child–parent implications and alternating XOR/OR label gates for dependency circuits. Layered admission clauses are $\neg x_i \vee \neg x_{i+1} \vee x_{i+3}$ in cyclic shuffled order, with seeded XOR/AND/OR read-once label trees. The domain sizes are respectively 81/729/6,561 and 80/680/5,984. A registered action $A = 1$ and $h = H(U) \wedge (A \vee \neg f(U))$ realize label f on domain H by disabling A. From the admissible all-zero reference, the first two shuffled slots become missing and the third flips to one. Budgets are $(2, 3), (3, 3), (3, 4)$; profiles $q = 1, 2$ protect every fourth shuffled position starting at the first. Seeds vary across profiles, so these compare whole queries; separate controls hold circuits fixed. All inputs precede the main run. A disjoint four-root development grid checks implementations against exhaustive record search.

Each case uses three repetitions. Worker time includes scientific imports and certificate handling; Bitset rebuilds its table. Symbolic/Bitset output caps are 4/2 MiB. Job size counts member bytes before outer ZIP compression. The cost tables label the independent checker as Reader.

Table 6: Certified costs: per-column medians over three repetitions, then six cases. Total aggregates per-run producer-plus-checker time and can differ from the sum of displayed medians.

Roots	Route	Producer (s)	Reader (s)	Total (s)	Job (KiB)
8	General	0.0847	0.1333	0.2218	453.9
8	Normal	0.0778	0.1286	0.2083	433.8
8	Bitset	0.0007	0.0011	0.0018	5.7
12	General	0.1857	0.3384	0.5226	1079.7
12	Normal	0.1815	0.3246	0.5041	1033.5
12	Bitset	0.0007	0.0153	0.0166	8.9
16	General	0.3868	0.8598	1.2728	2569.2
16	Normal	0.2951	0.7263	1.0212	1906.7
16	Bitset	0.0024	0.3099	0.3167	40.2

Frozen comparison and output caps. The cap study crosses those same eighteen cases with two symbolic routes, caps of 2/4/8/16 MiB and three repetitions (432 attempts). Route/cap order is balanced within repetition; there is no outcome-based instance filtering. Each producer and checker receives 8 seconds and 256 MiB RSS, with 2,000 solver calls for symbolic production. The 100 ms parent watchdog checks wall time, RSS and stored output; a terminal producer result above the cap is rejected before checker admission. The eighteen rejected attempts all occur at 2 MiB: fourteen already have complete producer outputs and four are killed during production. The other 414 attempts certify. A case completes only if all three repetitions certify.

Time ratios use paired producer-plus-checker medians for cases certified by both routes in all three repetitions. Coverage includes every attempt.

Table 7: Output-cap sensitivity on all eighteen cases.

Cap (MiB)	General	Normal	Normal/General time	Normal faster
2	14/18	16/18	1.003	4/14
4	18/18	18/18	1.006	7/18
8	18/18	18/18	1.000	9/18
16	18/18	18/18	0.996	10/18

Query/CNF/JSON retention accounts for the low-cap completion difference; raw proof length alone does not. Complete attempt ledgers and earlier post-hoc check results are retained in the artifact.

Explicit-table baseline and independent certificate. Each Bitset invocation rebuilds the complete factual and intervened truth masks from the original SCM using bitwise gate evaluation. It searches within the $\sum_{i=0}^{a} \binom{|E|}{i} 2^i$ allowed-record ball in cost order, stopping at the first successful record rather than visiting the full ternary cube. A certificate stores both complete indexed label masks and the answer/witness. The independent checker uses a different, scalar two-execution interpreter on every binary root assignment, checking domain membership, labels and table completeness. It then reconstructs distance support with binary counting planes and checks the witness and every lower-cost record; unflippability requires the whole allowed ball. It imports neither the producer nor a SAT solver. Different certificate formats thus preserve the scientific obligations, rather than forcing the table method to emit SAT logs.

All 54 Bitset attempts certify at 2 MiB under the same time/RSS limits. Jobs occupy about 5–43 kB; indexed truth masks use at most 16 KiB before JSON. Table 8 separates disjoint phases. Symbolic encoding, solver/proof and serialization counters are nested and retained separately.

Build constructs the complete table and Reader table independently reconstructs it. Optimality includes coordinate masks and lower-cost checks. Parent total includes startup and monitoring. Columns aggregate separately.

Table 8: Bitset phases in milliseconds, aggregated across six case medians.

Roots	Build	Search	Write	Reader table	Optimality	Parent total
8	0.030	0.198	0.174	0.671	0.227	211.056
12	0.096	0.217	0.189	11.874	2.807	211.721
16	1.168	0.676	0.291	243.026	60.590	474.889

At 4 MiB, Bitset is faster on sixteen cases in both worker and parent time. Its median worker ratios to General/Normal are 0.0293/0.0306; parent ratios are 0.438/0.439. Process startup and 100 ms polling explain much of this gap. Both symbolic routes are faster on the two sixteen-root layered cases, where Bitset pays for complete table reconstruction.

Implementation and environment. Sequential measurement batches ran on Apple M4 (10 CPU cores, 16 GiB RAM), macOS 26.5.1 build 25F80, Python 3.12.7, with one producer or checker process at a time. Symbolic production uses Glucose3 through `python-sat` 1.9.dev15 (Ignatiev et al., 2018) with native DRUP output; `psutil` 5.9.0 supports monitoring. The frozen independent Python checker reconstructs CNFs from the SCM and checks assignments and watched-literal RUP propagation without SAT calls; it ignores deletion hints. The supplement includes its Normal-route specialization and unchanged RUP kernel. Bitset production/checking uses standard-library Python. Worker wall starts before importing each scientific module and includes certificate writing/checking. Parent wall additionally includes process startup and watchdog polling; outer ZIP compression is archival overhead. Timing batches do not overlap other project tests, but the host OS is not isolated. The three repeats describe host variability, not independent tasks.

G.2 Persistent-core controls and all-optimal classification

Cost-layer enumeration retains every optimum for the eighteen inputs: two initially definite and sixteen ambiguous. All have empty cores; only two require mixed edits. Full radius and edit-type distributions accompany the unchanged cost records in the artifact.

The additional controls use the original eight-root dependency and layered SCMs. Each crosses four registered conditions: complete zero center with $a = b$ equal to the nearest opposite-world distance; the same center with $(a, b) = (1, 4)$; the original two-missing/one-wrong center with $a = b = 3$; and that damaged center with $(a, b) = (2, 3)$ and only missing slots editable. These are structure-selected boundary controls, holding each domain and label map fixed. All routes use 4 MiB output, 8 seconds and 256 MiB per stage, with three fresh-process repetitions and balanced route order.

Here D denotes erasure, V value assignment, and "or" alternative optimal types. Infinity means bounded unflippability. Calls/cubes are three-repeat General/Normal medians; Pass includes all routes and repetitions.

Table 9: Core controls and optimal edit types.

Family / control	Core labels	ρ^*	Optimal edits	Calls G/N	Cubes G/N	Pass
Dependency / clean, matched	0	1	D or V	26/26	2/2	9/9
Dependency / clean, slack	$0, 1$	∞	–	26/26	2/2	9/9
Dependency / damaged, matched	$\varnothing$	2	D+V or V	109/99	10/9	9/9
Dependency / damaged, fill only	$0, 1$	∞	–	28/28	2/2	9/9
Layered / clean, matched	0	1	D	26/26	2/2	9/9
Layered / clean, slack	$0, 1$	∞	–	26/26	2/2	9/9
Layered / damaged, matched	$\varnothing$	1	D	27/27	2/2	9/9
Layered / damaged, fill only	$0, 1$	∞	–	28/28	2/2	9/9

In the dependency empty-core control, Normal's total worker time is 185 ms versus 199 ms for General, with one fewer cube. Across the other controls, worker differences have both signs, while

symbolic cube and solver-call counts agree. Complete bitset construction and independent checks take 1.3–1.7 ms per control in worker medians; process startup dominates their invocation time. The artifact separates search-solver counters, total producer/checker costs, startup, counterexample counts, cube sizes and certificate bytes for every run.

G.3 Joint-width and budget profiles

Table 10: Peak reachable distance profiles at fixed length and edit budget. All twelve program queries are certified.

Joint width	$b = 2$	$b = 3$	$b = 4$	$b = 6$
5	63	111	138	152
7	103	188	252	280
11	122	228	331	569

The program family fixes $r = 24$, $a = 4$, missing positions eight and sixteen, and zeros elsewhere. The domain disallows adjacent ones and the label is one exactly when the number of ones is congruent to one modulo $p \in \{2, 3, 5\}$. Unreachable states and equal suffix functions are merged after completing all levels, giving actual joint widths 5, 7 and 11. Crossing these with $b \in \{2, 3, 4, 6\}$ gives the twelve cells of Table 10. Programs are supplied inputs; all roots affect the domain or label.

The producer stores the minimum edit cost for every retained profile at every layer. A separate checker reconstructs incoming transitions and checks complete Bellman closure and minima; exact-cost propagation independently checks the center and final witness. Three fresh processes per cell have 8 seconds, 256 MiB and 8 MiB output per stage. All attempts certify. Small exhaustive world/record controls and targeted certificate tampering check the implementations.

P/R worker times include imports, parsing and certificate handling. Parent time adds startup and 20 ms monitoring. RSS includes interpreter and stored layers; Proof counts gzip bytes before outer ZIP compression.

Table 11: Distance-profile costs for the twelve cells; entries are three-repeat medians.

w	b	Total profiles	P/R (ms)	Parent (ms)	RSS P/R (MiB)	Proof (KiB)
5	2	1,256	6.3/11.0	97.9	15.03/14.44	1.11
5	3	1,983	8.4/15.7	97.6	15.17/14.64	1.79
5	4	2,339	9.4/16.9	98.3	15.28/14.73	2.13
5	6	2,488	9.6/17.5	97.6	15.45/14.73	2.29
7	2	1,904	10.0/21.3	98.2	15.33/14.81	2.31
7	3	3,278	16.1/33.5	124.8	15.75/15.14	3.99
7	4	4,124	19.6/39.4	122.1	16.48/15.33	5.14
7	6	4,416	20.1/41.2	123.0	16.47/15.47	5.49
11	2	2,123	17.7/36.2	120.4	15.66/15.06	4.22
11	3	3,814	33.7/60.9	173.0	16.23/15.83	7.89
11	4	5,402	45.0/80.8	197.5	17.36/16.39	11.78
11	6	6,830	52.2/96.4	223.1	17.66/17.05	15.48

This implementation retains all layers for witness recovery and independent checking; its measured memory is not the two-layer working-space bound. Machine and Python environment match the other current measurements, with formal batches run sequentially. Per-layer profile counts, transition counts and complete attempt ledgers are retained as data.

G.4 Independent edit- and filter-budget diagnostic

This diagnostic tests the edit-budget branch of the state bound on supplied programs. The domain consists of length-24 strings with no adjacent ones; the label is one exactly when the ones-count

is congruent to one modulo $p \in \{2, 3, 5\}$. Completed joint widths are 5, 7 and 11. We fix two centers before measurement: zeros with slots 8 and 16 missing, and the all-one record. All slots are editable. The former crosses $a \in \{1, 2, 4\}$ with $b \in \{2, 4, 12, 14\}$; the latter uses the same a values and $b \in \{12, 14\}$, which yield nonempty initial support. These are 54 configurations.

Three rotated repetitions use the unchanged producer and independent checker, with eight seconds and 256 MiB RSS per stage and an eight-MiB total job-output cap. The run uses Python 3.12.7 on macOS 26.5.1 (arm64), with psutil 5.9.0 for resource monitoring. The registry, inputs, all 162 attempts and the complete 54-row parameter table are included in the reproduction package. All attempts complete and certify: 17 configurations have finite radii (six at one, eight at two, three at four), and 37 are unflippable within their edit budgets. Of the 45 configurations with $2A+1 < B+2$, eight have finite radii and 37 have no nonempty report change within budget.

Table 12: Profile and certification costs at the all-one center, $r = 24$, $w = 7$.

b	a	ν	ρ^*	Peak	Total	P (ms)	C (ms)	KiB
12	1	3	> 1	7	158	3.4	3.0	3.5
12	2	5	2	29	609	4.5	6.3	12.9
12	4	9	2	224	4109	28.2	36.2	85.6
14	1	3	> 1	7	159	3.3	2.9	3.6
14	2	5	> 2	29	613	4.5	6.2	12.9
14	4	9	> 4	224	4136	28.2	36.1	86.1

The all-one center has initial report $\{0\}$ at $b = 12$ and $\{0, 1\}$ at $b = 14$. In Table 12, Peak is the maximum number of retained profiles at a layer, Total sums over layers, P/C are median producer/checker worker times, and KiB measures the raw certificate. Across all configurations, peak profiles range from 7 to 634 and raw certificates from 2.9 to 173.4 KiB. Per-case median P/C times span 3.12–53.28 and 2.57–97.05 ms; including process startup and monitoring gives 26.40–97.70 and 26.41–124.93 ms, respectively. Increasing the edit budget can expand profiles, whereas changing the filter budget can change reachability at the same peak count. Both state bounds analyze the same recurrence; the measurements separate these parameter effects from process and certificate costs.

H File-Execution Evidence and Audit Queries

H.1 Physical interpretation and query domains

The registry fixes initial-state bits and action gates as atomic evidence slots, their order and unit charges. Every compared method receives the same $\mathcal{H}, y, o_0, E, a, b$; factual roots and corruption locations supply evaluation references. Controlled erasure and replacement give the known initial bound $b_0 = 2$ in the radius study. Its fixed $b = 2$ filter is tested under two additional edits; the separately configured $b = 4$ filter covers their cumulative bound. The query uses the resulting record; edit histories are retained for checking. Fills are unverified record writes charged by changed slot. Field protection fixes the current value, and schema, permissions and charges remain fixed within each comparison. These are the registered units of the controlled study.

Archive protects two original reports and registers backup locations; harm means a protected object has no recoverable original copy. Its four fixed plans cover all registered intervention contexts, with two half-score worlds sharing a joint-deletion mechanism. Rollback tracks live, checkpoint and trial bytes across four plans sharing the capture/trial/restore/cleanup structure. Each context starts from its registered snapshot; interventions change later tool effects and returns within the fixed sequence. The artifact retains the measured world-outcome tables.

Rollback records 1,280 post-step byte states. Recovery checks actually read the registered paths and compare original bytes, independently of model judgments. Intervention results reconstruct the minimal cause family under the registered safe-action domain and no freezes. For the two restoration histories in Figure 2, the checkpoint is left available when it is stale and is removed by cleanup in the captured-original history. Thus failure to restore in the latter can lose the last recoverable original.

H.2 PLAN PROVENANCE AND RESPONSE-FUNCTION UNITS

Archive tasks register four initial archive states: neither valid, only the first valid, both valid, and the first stale with the second absent. The goal is to preserve both originals and clear the inbox via list/read/copy/remove/finish. Rollback uses configuration, document, catalog and script files with an original, distinct candidate and stale checkpoint; its goal is capture, trial, restore and checkpoint cleanup. Two fresh-context same-team Codex sessions each received four tasks, one attempt per task, at most 20 calls including finish and four mutation attempts including failures. Every task is retained without retry or replacement. All natural executions had measured harm zero; interventions replay the frozen calls. Recovered session metadata records requests to OpenAI for `gpt-6-astra` with `ultra` reasoning, using Codex CLI `0.153.4` for archive and `0.154.0-alpha.6.2` for rollback. All 28 archive and 40 rollback task requests match the frozen journals. The design-code supplement's `provenance/ACTOR_PROVENANCE_SANITIZED.json` records this retrospective extraction of requested configuration. Temperature, `top_p`, random seed, and the backend-served snapshot were not recorded.

Archive t00/t03 share a six-root response (15 harmful worlds each); t01 has five roots, three relevant (four harmful worlds); t02 is constant-safe (zero of 16). All four rollback plans share a five-root response (nine of 32). Removing irrelevant slots leaves three nonconstant functions and one constant. The 304 contexts count assignments within plans. Radius rows additionally specify candidate, harmful reference, missing/flipped coordinates and profile. Task-indexed deduplication retains plan identity.

H.3 EXECUTION-DERIVED TARGET RADII AND PROTECTION CHOICES

Measured recovery supplies the harmful domain and singleton-prevention labels. For each harmful assignment and candidate, erasing one root and reversing another gives a center at distance two from its factual reference. With $a = 2$, the profiles use $b = 2$ on all slots, $b = 2$ on action gates, and $b = 4$ on all slots.

Independent breadth-first traversal checks every target minimum and attaining record for 8,100 task-indexed queries and 20,160 registry rows. The query key is (panel, task, candidate, profile, center); it retains identical response functions in different plans. Across all profiles, 1,536 of 2,704 rows (860 of 1,332 query keys) have an opposite report at higher cost than their ambiguity-only cheapest change. Restricting to $b = 2$ gives 1,536 of 2,696 rows (860 of 1,328 keys). The four additional $b = 4$ keys, eight rows, have center $1110\mathtt{M}$ for candidate 2 in the four rollback plans; only ambiguity is reachable within $a = 2$. In Table 13, N_A counts definite rows whose minimum nonempty changes only reach ambiguity; N_+ reach the opposite singleton later. U_D, U_A count unflippable definite/ambiguous rows; $U_{D,\emptyset}$ still reach empty support. Each archive/rollback profile has 3,840/2,880 rows.

Table 13: Target reachability and the meaning of bounded unflippability.

Panel	Profile	N_A	N_+	U_D	U_A	$U_{D,\emptyset}$
Archive	$b = 2$, all	816	384	2,040	0	2,040
Archive	$b = 2$, gates	816	384	2,040	0	2,040
Archive	$b = 4$, all	0	0	2,040	1,080	0
Rollback	$b = 2$, all	532	388	728	0	728
Rollback	$b = 2$, gates	532	380	728	0	728
Rollback	$b = 4$, all	8	0	720	1,360	0

The retained mixed-edit witness is $001\mathtt{M}1$ for trial, with initial checkpoint validity protected. Both singleton targets have radius two; $0\mathtt{M}0\mathtt{M}1$ supports only 00011, label zero, excluding its label-one factual reference 00100. All twenty necessarily-mixed registry rows share this candidate/center pattern across four plans and five factual references. At $b = 4$, every attainable resolution instead has a factual-label pure-deletion optimum. The registered inputs and certification ledgers for the earlier 81-query symbolic sample and separate mixed witness are retained as data.

A uniform rule using the observed report. The post-hoc comparison protects two action gates in all initially definite gate-editable queries: 4,552 rows, 2,048 task-indexed queries. For $G_b(o_0) = \{\ell\}$, put $r_o = \rho_{\{1-\ell\}}$, $r_g = \rho^*$ and $r_c = \min_{T \neq \{\ell\}} \rho_T$, including the empty target. Target maximizes (r_o, r_c, r_g) lexicographically; Global maximizes (r_g, r_o, r_c), using the opposite target as its first tiebreak. Direct uses Target's order and requires protecting the queried action's gate. All maximizers are retained; bounded-unreachable values rank above a. Policies use the record, its report and domain/labels; the factual root serves only containment and reference checks.

Table 14 gives blocking outcomes over all 4,552 rows (2,048 task-indexed queries); 1,200 initially admit the opposite target. Target's primary optimum weakly dominates Global's and Direct's by construction. The strict difference from Global is eight rows at two centers across four rollback plans. The 24 remaining cases are shared failures. Target's certainty radius is larger/equal/smaller than Global's in 4/4,548/0 rows, and than Direct's in 828/3,724/0. Depending on retained ties, empty support remains reachable in 2,048–2,516 Target rows, 2,052–2,520 Global rows and 2,604–3,016 Direct rows. These are exhaustive comparisons of registered field choices and their report targets.

A complete protection example. For the restore query, slots are (checkpoint validity, capture, trial, restore, cleanup). Measured harm and singleton prevention give $\mathcal{H}_\ell = \{w \in \mathcal{H} : y(w) = \ell\}$:

$$\mathcal{H}_1 = \{00010, 00011\},$$
$$\mathcal{H}_0 = \{00100, 00101, 00110, 00111, 01101, 10101, 11101\}.$$

At $o_0 = \texttt{M1010}$, $a = b = 2$, with the checkpoint slot fixed, the initial support is $\{00010\}$. Exhausting the action-slot edit ball gives these inclusion-minimal changed-coordinate sets and witnesses:

Target	Minimal successful edit sets	Witnesses
$\{0\}$	$\{\text{trial}\}$, $\{\text{restore}, \text{cleanup}\}$	M1110, M1001
$\{0,1\}$	$\{\text{capture}\}$	M0010
$\varnothing$	$\{\text{trial}\}$, $\{\text{restore}\}$, $\{\text{cleanup}\}$	M1M10, M1000, M101M.

Every successful edit set contains a listed minimal set. By Appendix B.3, protection must hit each set: blocking the opposite report requires trial and either restore or cleanup; blocking empty support requires all three. These form the unique minimum joint protection set, while capture still permits ambiguity at cost one. Blocking ambiguity too requires all four. This derives Table 4; its opposite witness M1001 supports $\{01101, 11101\}$. The same objective conflict occurs at 1M010; the all-slot profile has the corresponding eight conflicts at protection cost three.

Table 14: Two-field protection in the fixed-$b = 2$ file panels.

Rule	Opposite target blocked	Newly blocked
Target	4,528/4,552	1,176/1,200
Global	4,520/4,552	1,168/1,200
Direct	4,428/4,552	1,076/1,200

H.4 Supplementary conditional decisions

This study uses the same measured labels to compare decisions under a prespecified loss, complementing the target-radius protection analysis. Harm h, candidate score s_i, threshold label y_i and action loss $\ell(a, y_i)$ are distinct objects. Investigate escalates candidate review; defer leaves that query unresolved; close ends only the candidate's high-score query. At the primary strict threshold $1/2$, investigate/defer/close losses are the prespecified audit losses $(4, 1, 0)$ for label zero and $(1, 3, 8)$ for label one. A rule minimizes worst-case loss over its supported labels (Manski, 1999), breaking ties by defer/investigate/close.

Full retains all compatible harmful worlds. Point selects a world by explanation cost then lexicographic order. Ties retains every minimum-cost world. Membership uses full's worlds but retains only cause membership: if $m(w)$ candidates occur in at least one minimal cause, a member contributes $\{1/j : 1 \leq j \leq m(w)\}$ and a nonmember $\{0\}$, united across worlds. Always-defer ignores the evidence. Execution-derived factual references enter only the evaluation; the rules use their evidence-derived supports.

Separate panels enumerate every single-slot erasure and replacement at budget one for eligible harmful contexts: 1,560 archive and 1,440 rollback queries.

Against ties, archive has 24 wins, 1,500 ties and 36 losses, a net saving of 36 loss units, all contributed by its 48 true half-score cells (four candidate–world pairs). Rollback has 48 wins, 1,320 ties and 72 losses, net 88: half-score queries contribute 32 and score-one queries 56. Full and ties agree on all 780 archive and 720 rollback erasures; differences arise from replacements. The artifact retains task-equal and half-score strata.

Table 15: Primary decisions at strict threshold $1/2$. FD counts incorrect definite labels against the intervention reference.

	Archive: 1,560 queries			Rollback: 1,440 queries		
Rule	Mean loss	FD	Defer	Mean loss	FD	Defer
Full	0.395	0	156	0.533	0	248
Point	0.499	72	0	0.722	112	0
Ties	0.418	24	96	0.594	48	128
Membership	0.433	0	216	0.928	0	608
Always defer	1.482	N/A	1560	1.556	N/A	1440

Threshold and loss sensitivity. At threshold $1/3$, gain over membership vanishes in both panels; rollback's gain over ties remains 80. Holding all else fixed, a post-hoc archive defer-cost multiplier α reverses the full/ties mean ordering exactly on $(14/11, 4/3]$ in the reported cost scan; the endpoint $4/3$ includes the registered tie action. For rollback, changing only the high-label close cost to c gives, for $c \geq 3$,

$$L_F = \frac{8}{15}, \qquad L_T = \frac{83 + 3c}{180}, \qquad L_T - L_F = \frac{3c - 13}{180}. \tag{23}$$

Ties is better for $3 \leq c < 13/3$, equal at $13/3$, and full better above it. The primary $c = 8$ result is unchanged. Below three, selected actions change. These comparisons measure information value under the stipulated audit losses.

H.5 Decision structure and threshold information

If the factual world is compatible, every full-set definite label agrees with it. Full's zero false-definite count follows from this containment property. Membership has a superset of full's label support. For any fixed true label in full's support, the common rule cannot improve on full by replacing a pure support with $\{0, 1\}$: pure supports select their label-optimal action, while mixed supports use the same minimax action. This proves its structural nonadvantage under this contract; the observed strict gains are empirical.

For the registered inverse-cardinality score and empty freeze domain $F = \varnothing$,

$$s_i(w) > 1/2 \quad \Longleftrightarrow \quad \{i\} \text{ is a minimal cause} \quad \Longleftrightarrow \quad \text{disabling } i \text{ alone prevents harm.} \tag{24}$$

Indeed a positive score is $1/k$, so exceeding $1/2$ requires $k = 1$; factual harm makes the empty intervention unsuccessful. More generally, strict $1/3$ permits causes of size at most two. Threshold-specific queries reproduce both panels' full/point/ties decisions without exact scores. At the primary threshold, outcome-table reads fall from 76 to 66 (archive) and 92 to 76 (rollback); at $1/3$ or jointly they do not fall. Full timing records show no uniform speedup in the measured inference pipeline. These outputs differ from complete minimal families, exact scores or radius certificates.