CARTER CERTIFIES PARTIAL TAXONOMY REPAIRS: COMPLETE DECISIONS AND WITNESSES UNDER NOISY FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Repairing a taxonomy requires deciding which parent relations to change and which to leave unresolved. With noisy feedback, agreement among surviving trees can increase after the target has been discarded. We formalize this failure as *confident erasure* and introduce CARTER, an exact interface that turns a calibrated family of primary-parent trees into complete edge certificates and opposing witnesses. A session-calibrated disagreement budget preserves target retention and all certified edits at every prefix with marginal probability at least $1 - \alpha_s - \delta_h$, under candidate inclusion and exchangeable sessions from a frozen pipeline. Errors within a session may be dependent. Edge-additive disagreement reduces certification to constrained arborescence optimization, recovering every edge statement shared by the retained family without enumerating trees. An optional shared determinant circuit computes exact one-step query values. On a sealed structural holdout of 224 taxonomies, six-source macro-average whole-tree coverage is 91.5–93.5% at a 90% target, versus 34.6–41.2% for edge-pooled calibration. A separate development simulation adds 2.095 final certified edits per session over no feedback, with lifetime wrong edits in 2/148 sessions. These results connect structured uncertainty to inspectable partial repairs, with explicit witnesses for deferred decisions.

1 INTRODUCTION

Taxonomies organize concepts into parent–child hierarchies. MeSH supports biomedical indexing, while AGROVOC organizes agricultural knowledge (U.S. National Library of Medicine, 2026; Food and Agriculture Organization of the United Nations, 2026). Maintaining such resources involves deciding which parent changes to commit or defer. We study partial repair of a designated primary-parent tree, certifying additions and deletions while retaining ambiguous choices. This target remains explicit when the source permits multiple parents.

A wrong answer can increase apparent certainty while removing the target. In Figure 1(a,b), Dog's observed parent is Bird and its target parent is Mammal. An incorrect "Bird" answer makes hard filtering discard the target. The surviving family agrees on new edge statements; target loss can precede any wrong edit. We call this failure *confident erasure*: greater agreement need not mean better evidence.

Preventing this failure requires representing structural uncertainty and updating it under unreliable feedback. Chen et al. (2021) reconcile language-model parent scores into one consistent tree, while Zhang et al. (2025) provide compact representations of prediction sets. These approaches support tree construction and uncertainty representation; repair additionally requires updating a complete-tree family under fallible answers and deciding which edge statements hold throughout it. Edge-wise coverage alone does not ensure whole-tree coverage at the same level. Moreover, the family can be exponentially large, and combining individually possible edges can violate its disagreement budget. Selecting one best tree cannot resolve these family-wide decisions.

To address this gap, we introduce CARTER (Calibrated Adaptive Refinement of Taxonomies with Error-robust Responses). It retains initially plausible trees unless their accumulated disagreement exceeds a calibrated tolerance, and commits only edits supported by every retained tree. Whole-tree calibration forms the initial family; complete sessions from a frozen pipeline calibrate the feedback

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.

tolerance. Because disagreement cannot decrease, covering the terminal count protects every earlier prefix. Edge-additive disagreement enables exact constrained arborescence tests, avoiding tree enumeration; opposing feasible trees explain unresolved edges. With matched session scores and endpoint conventions, binary conformal risk control (CRC) on budget-exceedance loss selects the same budget on the full integer grid (Angelopoulos et al., 2024); shared transcripts and solvers then yield identical states and certificates (Appendix B.6). CARTER connects this calibration to complete structured repair decisions: every universally supported edge statement is returned, and each unresolved edge has two feasible trees explaining the disagreement.

Our contribution is a *complete structured certificate interface* with three properties. **Session-wide validity:** a covered target makes all edits sound simultaneously across the interaction. **Exact decisions and witnesses:** at most $|D_0| + 1$ arborescence optimizations recover all mandatory, possible, and impossible edges of the budgeted family, including feasible counterexamples for every unresolved edge. **Shared query computation:** one forward–reverse arithmetic circuit evaluates all admissible binary questions for an optional one-step policy. A sealed structural benchmark measures coverage and editing output; programmed-feedback studies measure retention, wrong edits, and calibration variability. Together with enumeration and reference-solver checks, these experiments test the statistical and computational parts of the interface.

2 CERTIFIED REPAIR WITH BUDGETED VERSION SPACES

2.1 OVERVIEW

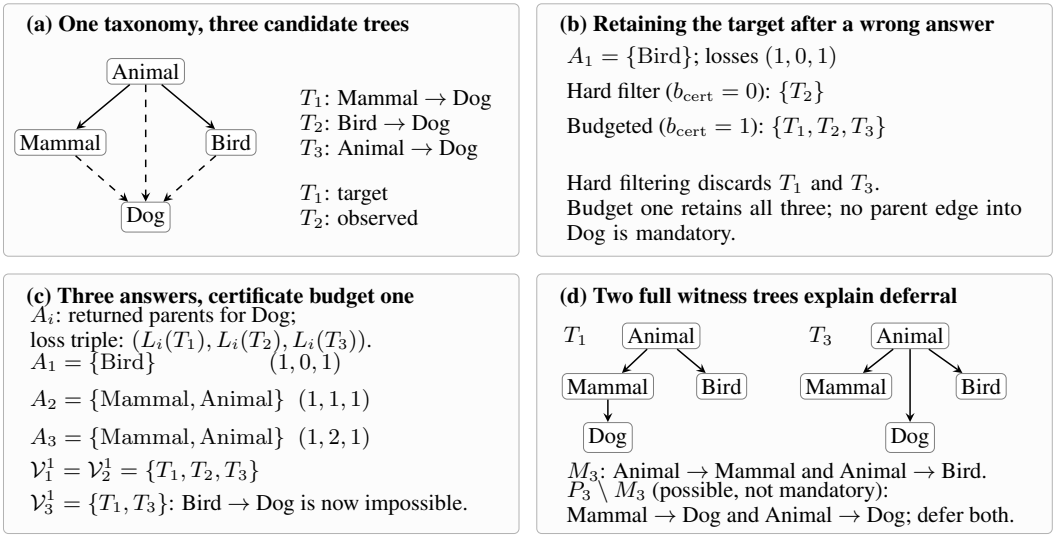


Figure 1: Schematic four-node running example of partial repair. (a) Each tree uses both solid edges and one parent edge into Dog. (b) A wrong answer removes the target under hard filtering, but not with $b_{\text{cert}} = 1$. (c) Loss triples give the values $(L_i(T_1), L_i(T_2), L_i(T_3))$. (d) Surviving witnesses disagree on Dog's replacement parent: delete Bird→Dog but defer both alternatives. M_3 and P_3 denote edges present in every and at least one retained tree, respectively. States and witnesses match enumeration and both certificate implementations.

CARTER takes an observed graph and scored candidate edges. Labeled taxonomies calibrate its initial tree family; disjoint complete feedback sessions calibrate the disagreement tolerance b_{cert}.

Online, a frozen policy selects a query; the response is an acceptable-parent set or an abstention, with abstention adding no disagreement. A tree remains in the *version space* if it belongs to the initial family and contradicts at most b_{cert} answers so far. When used, the separate utility budget b_{qry} governs query selection and is fixed before session calibration; it does not replace the certificate budget.

Exact tests certify additions present in every retained tree and deletions present in none. Opposing trees explain unresolved choices; an empty family triggers conflict and no edits. Figure 1(c,d) deletes Bird→Dog while deferring its replacement between Mammal and Animal.

2.2 Problem setup and repair states

Inputs and target. We model a complete taxonomy as a rooted spanning tree, or *arborescence*, assigning one parent to each non-root concept. Let V be its nodes, r the designated root, E_{obs} the observed edges, and $C \subseteq \{(u, v) \in V \times V : u \neq v,\ v \neq r\}$ a frozen candidate-edge universe. Write $\mathcal{T}(G, r)$ for the arborescences of graph G rooted at r. The *canonical target* $T^\star$ is a predeclared primary-parent tree, never inferred from deployment feedback. For source graphs with multiple parents, the benchmark selects a parent by shortest root depth, then by node identifier; Appendix A.2 specifies the projection. Correctness means agreement with this frozen tree. Other source-supported parents retain their separate semantic status. Our theorems assume $T^\star \in \mathcal{T}((V, C), r)$ almost surely and $0 < \alpha_s, \delta_h$ with $\alpha_s + \delta_h < 1$. Appendix A.1 tabulates the notation.

Initial tree family. Smaller frozen nonconformity scores $s(e)$ indicate more compatible edges; we use adaptive prediction sets (Romano et al., 2020) over each child's candidates. The complete-tree score $S(T) = \max_{e \in E(T)} s(e)$ takes the worst score among tree edges $E(T)$. Let q be its split-conformal threshold. A graph-only feasibility closure $q' \geq q$ raises it only until every node is reachable from r through $D_0 = \{e \in C : s(e) \leq q'\}$, using scores and graph structure, never the new instance's target labels. Then $H_0 = (V, D_0)$ induces the initial family $\mathcal{V}_0 = \mathcal{T}(H_0, r)$, exactly the tree-score level set at q'. Under exchangeable whole-taxonomy calibration and deployment scores, split conformal calibration (Vovk et al., 2005; Lei et al., 2018) and this set-enlarging closure give

$$\Pr\{T^\star \notin \mathcal{V}_0\} \leq \alpha_s. \tag{1}$$

Theorem 1 holds for any covered set; Theorems 2–3 require this graph-induced form (Appendix A).

Feedback and retained trees. At paid round s, an adaptive policy queries a non-root child v_s and receives a nonempty acceptable-parent set A_s or an abstention $\perp$. Encoding abstention as $A_s = \{p : (p, v_s) \in D_0\}$ consumes a query but adds no disagreement. For certificate budget $b_{\text{cert}} \in \mathbb{Z}_{\geq 0}$, define

$$L_t(T) = \sum_{s=1}^{t} \mathbf{1}\{\text{pa}_T(v_s) \notin A_s\}, \qquad \mathcal{V}_t^{b_{\text{cert}}} = \{T \in \mathcal{V}_0 : L_t(T) \leq b_{\text{cert}}\}, \tag{2}$$

where $\text{pa}_T(v)$ is v's parent in T. The *budgeted version space* retains trees contradicting at most b_{cert} answers.

Repair decisions. An empty $\mathcal{V}_t^{b_{\text{cert}}}$ triggers conflict and no certificates. Otherwise, define

$$M_t = \bigcap_{T \in \mathcal{V}_t^{b_{\text{cert}}}} E(T), \quad P_t = \bigcup_{T \in \mathcal{V}_t^{b_{\text{cert}}}} E(T), \quad I_t = C \setminus P_t, \tag{3}$$

where *mandatory* edges M_t occur in every retained tree, *possible* edges P_t in at least one, and *impossible* edges I_t in none. Thus P_t includes M_t; $P_t \setminus M_t$ remains unresolved. Certified additions are $M_t \setminus E_{\text{obs}}$ and deletions $E_{\text{obs}} \setminus P_t$ (Figure 1(d)). Deletions include observed edges outside C, which no candidate tree can contain.

Definition 1 (confident erasure). For a nonempty tree family $\mathcal{W}$, collect its unanimous edge statements as

$$\Gamma(\mathcal{W}) = \{+e : e \in \cap_{T \in \mathcal{W}} E(T)\} \cup \{-e : e \in C \setminus \cup_{T \in \mathcal{W}} E(T)\}.$$

An update to a nonempty family $\mathcal{W}'$ causes *confident erasure* when $T^\star \in \mathcal{W}$, $T^\star \notin \mathcal{W}'$, and $\Gamma(\mathcal{W}') \setminus \Gamma(\mathcal{W}) \neq \varnothing$: it removes the target while creating new unanimous statements. Hard filtering can do this after one wrong answer; the budgeted update retains an initially covered target until its tolerance is exceeded. Confident erasure is one form of target loss; Theorem 1 protects against target loss itself, including loss with no new unanimous statement. Empty-family conflicts are reported separately.

2.3 CALIBRATING THE FEEDBACK BUDGET

Calibrating complete sessions matches the unit protected by the guarantee. Run the frozen pipeline on n reliability-calibration sessions and count target-parent exclusions in each. An order statistic of these counts sets the tolerance:

$$k = \lceil (n+1)(1-\delta_h) \rceil, \qquad b_{\mathrm{cert}} = \begin{cases} Y_{(k)}, & k \le n, \\ H_{\max}, & k = n+1, \end{cases} \tag{4}$$

where $Y_j = L_{\tau_j}(T_j^\star)$ counts answers excluding the canonical parent in calibration session j, τ_j is its number of paid questions, $Y_{(k)}$ is the k-th smallest of $Y_1, \ldots, Y_n$, and $H_{\max}$ is the prespecified maximum number of questions per session. In the second case, no session can exceed the budget. Split conformal calibration (Vovk et al., 2005; Lei et al., 2018) uses one score per session, rather than per edge, to cover the next exchangeable session with probability at least $1 - \delta_h$. Dependence among answers within a session is permitted. Calibration has limited resolution: at $\delta_h = 0.05$, fewer than 19 sessions require the maximum budget. Feedback then leaves $\mathcal{V}_0$ unchanged, retaining its structural certificates but adding no feedback-induced ones. Appendix A gives a finite-population alternative for release-time auditing. The next subsection states the frozen-pipeline assumptions and connects this terminal budget to every interaction prefix.

2.4 EVERY-PREFIX CONTAINMENT GUARANTEE

Calibration protocol. Freeze the scorer and candidate rule on training data disjoint from both calibration streams. Structural calibration fixes the threshold rule. Conditional on that first stream and the resulting frozen procedure, assume that sessions in a disjoint reliability-calibration sample and the next deployment session, generated by the same frozen pipeline, are exchangeable. Before reliability calibration, freeze the query policy, its randomization and utility parameter b_{qry}, and the answer-aggregation protocol. Also fix a deterministic horizon $H_{\max}$ and a stopping rule that uses only information available so far, yielding $\tau \le H_{\max}$ paid questions. Section 2.3 then yields

$$\Pr\{L_\tau(T^\star) > b_{\mathrm{cert}}\} \le \delta_h. \tag{5}$$

Thus α_s controls initial structural miscoverage, whereas δ_h controls complete-session budget exceedance. Within-session errors may be dependent, persistent, or adaptively located.

A candidate failure, where C excludes $T^\star$, violates the theorem's inclusion assumption. Within that scope, the two controlled failures are $T^\star \notin \mathcal{V}_0$ and $L_\tau(T^\star) > b_{\mathrm{cert}}$. Monotonicity extends terminal budget coverage to every prefix; a union bound combines it with initial structural coverage.

Theorem 1 (lifetime all-edit safety). Over repeated draws of both calibration samples and the next exchangeable session, with probability at least $1 - \alpha_s - \delta_h$, $T^\star \in \mathcal{V}_t^{b_{\mathrm{cert}}}$ simultaneously for every $t = 0, \ldots, \tau$. On the same event, every certified addition and deletion at every round is correct relative to the canonical target tree.

Proof sketch. The two failure events require no independence. When neither failure occurs, $L_t(T^\star) \le L_\tau(T^\star) \le b_{\mathrm{cert}}$ for every prefix, so the target witnesses every robust set and every intersection or union statement is sound. All decisions follow from that one event, so no edge-wise or round-wise multiplicity correction is needed (Appendix A). With $\alpha_s = \delta_h = 0.05$, the 0.90 bound is marginal and session-level, jointly covering target retention and all certified edits at every prefix, rather than a posterior confidence assigned to each edge or round.

2.5 COMPLETE EDGE CERTIFICATES

The guarantee concerns an entire tree family. To compute its certificates without enumerating that family, charge each edge the answers it contradicts:

$$c_t(e) = \sum_{s \le t \,:\, v_s = v} \mathbf{1}\{u \notin A_s\}, \tag{6}$$

where $e = (u, v)$ and the sum accumulates all queries about child v, including repeated questions. Every arborescence selects one incoming edge per non-root concept, giving $L_t(T) = \sum_{e \in E(T)} c_t(e)$.

This edge-additive representation makes $\mathcal{V}_t^{b_{\text{cert}}}$ exactly the set of arborescences of H_0 with cost at most b_{cert}. Minimum-arborescence computation (Edmonds, 1967) therefore tests feasibility of this cost level set. All budget-feasible trees, including nonminimizers, contribute to its certificates.

Opposite-edge tests. First find a minimum-cost arborescence T_0 of $H_0 = (V, D_0)$. If $L_t(T_0) > b_{\text{cert}}$, report conflict. Otherwise, T_0 supplies one feasible witness for every edge decision. For an edge in T_0, forbid it and test whether an excluding tree remains within budget. For an edge in $D_0 \setminus E(T_0)$, force it and test for an including tree. A restricted problem with no arborescence has infinite optimum cost. A feasible opposite witness establishes that the edge is unresolved; otherwise all feasible trees agree with the baseline on the edge's presence. Edges in $C \setminus D_0$ are impossible and require no additional solve. Appendix A proves these tests; the released solver specifies deterministic tie handling and optional witness materialization.

Theorem 2 (complete robust certificate). If $\mathcal{V}_t^{b_{\text{cert}}} \neq \varnothing$, one baseline and one opposite-restriction solve per allowed edge recover exactly M_t, P_t, and I_t using at most $|D_0| + 1$ minimum-arborescence calls. Every edge in $P_t \setminus M_t$ has budget-feasible witnesses both containing and excluding it. Thus no additional edge statement can be added while remaining correct for every tree in $\mathcal{V}_t^{b_{\text{cert}}}$.

2.6 Exact one-step query selection via shared compilation

The optional query policy counts trees satisfying the utility budget b_{qry}. Each binary edge-membership question induces two successor states, evaluated with one fewer question remaining. Their Berlekamp weights combine disagreement-layer counts with the remaining answer-error allowance (Appendix A.5). The one-step rule minimizes the larger successor weight, with lexical ties. Its utility budget, tie-break rule, and empty-state fallback are frozen before reliability calibration; b_{cert} continues to govern certification.

For transcript ρ, compile additive edge costs $c_\rho(e)$ and define the tree polynomial $F(x) = \sum_{T \in \mathcal{V}_0} \prod_{e \in T} x_e$. Then

$$Z_\rho(z) = F(x)\big|_{x_e = z^{c_\rho(e)}}, \qquad M_{\rho,e}(z) = x_e \frac{\partial F}{\partial x_e}\bigg|_{x_f = z^{c_\rho(f)}}.$$

Coefficients through degree b_{qry} count all trees and those containing edge e, respectively, at each disagreement level. These counts determine the two successor weights. Here $M_{\rho,e}$ is a counting polynomial, distinct from the mandatory-edge set M_t.

One division-free determinant circuit and reverse accumulation share computation across edges. They specialize established tree-marginal and differentiation identities (Koo et al., 2007; Zmigrod et al., 2021; Baur & Strassen, 1983) to the truncated integer ring; avoiding division preserves validity even in the presence of zero divisors. Given the edge costs, the computation uses $O(d^4(b_{\text{qry}} + 1)^2 + |D_0|(b_{\text{qry}} + 1))$ arithmetic operations, $d = |V| - 1$; coefficient bit growth is excluded. Appendix A.5 provides the construction and full cost accounting.

3 Experiments

The evaluation separates the interface's two statistical stages. RQ1 and RQ2 test initial structural coverage and editing output on the sealed holdout, without feedback. RQ3 tests session-budget updates on separate development simulations. Computational and repeated-calibration diagnostics test exact implementation and variation across fitted feedback budgets.

3.1 Setup

Data. The sealed v5 study uses six source strata from five resources: Open English WordNet (OEWN) nouns and verbs, UNESCO, MeSH, AGROVOC, and the Australian and New Zealand Standard Research Classification (ANZSRC) (Open English WordNet Community, 2025; UNESCO, 2026; U.S. National Library of Medicine, 2026; Food and Agriculture Organization of the United Nations, 2026; Australian Bureau of Statistics, 2020). It excludes development nodes and contains 222 calibration and 224 node-disjoint test taxonomies; temporal transfer uses MeSH 2025–2026 and OEWN 2024–2025

(U.S. National Library of Medicine, 2025; Open English WordNet Community, 2024). Complete within-graph parent candidates include $T^\star$ (Appendices B.3 and B.9). The one-shot lock precedes target access but is an internal integrity record, not a third-party registration. Downstream checks recompute 35,840 transcripts from frozen scores without independently rerunning the upstream scorer; the earlier invalidated attempt remains documented in Appendix C.

Baselines. Point-output baselines comprise no-repair, local argmin, and maximum-score and minimum-edit arborescence decoders. Set-based certification compares edge pooling, per-child Bonferroni, and a finite-population learn-then-test adaptation (LTT; Angelopoulos et al., 2025). Feedback controls include hard filtering, plurality, fixed budgets, Dawid–Skene aggregation, and clean answers. Matched-interface CRC comparators isolate the role of budget calibration (Angelopoulos et al., 2024). Historical NR-OCP-inspired and raw output-CRC adapter records, including their reconstruction and applicability limits, are retained in Appendices B.7 and B.6. Operational definitions are in Appendices B and B.8.

Metrics. Whole-tree coverage indicates $T^\star$'s membership in the version space; graph-level wrong certification indicates any incorrect certified edit. The certified-edit count includes additions plus deletions, including incorrect ones; automation indicates at least one such edit. Both measure output volume. Correct-edit counts measure useful output, and wrong-certificate rates separately measure its safety. Point decoders instead report edge F1 and valid-tree rate. Table 1 averages five seeds within each graph, graphs within each source, then the six source means equally: a source-macro, not pooled-graph, estimand. Table 2 weights sessions equally. Its retained indicator requires target membership at every prefix; lifetime wrong requires an incorrect edit at any prefix; conflict requires an empty set at some prefix. Incremental is the paired final-edit increase over no feedback.

Implementation. Frozen edge scores combine word/character TF–IDF similarities and lexical features with features of the observed graph through a standardized logistic scorer. Domain-calibrated nonconformity uses $\alpha_s = 0.10$ structurally; canonical feedback uses $\alpha_s = \delta_h = 0.05$, $H_{\max} = 4$, and ambiguity-block queries, not the binary compiler. The checked parameter registry and configurations specify the exact arborescence implementation and separate query-audit settings (Appendix B).

Statistics. The sealed 20-test family comprises four corruption rates crossed with five contrasts: two edge-F1 comparisons, valid-tree rate, coverage, and wrong certification. Its source-stratified paired graph bootstrap and 100,000-draw, plus-one sign-flip tests retain within-graph seeds together; Holm adjustment covers this family (Holm, 1979). Sign flips require sign-exchangeable paired differences under the null, not randomized method assignment (Appendix B). Development intervals describe their stated settings; session and outer-fit resampling are distinguished below.

3.2 RQ1: Does whole-tree calibration preserve the target structure?

[Sealed holdout.] Whole-tree calibration retains substantially more target structures than edge pooling in this six-source evaluation. At $\alpha_s = 0.10$, CARTER has source-macro coverage of 0.915–0.935 and a wrong-certificate rate of at most 0.001 across four corruption rates, versus 0.346–0.412 coverage and 0.224–0.440 wrong certification for pooling (Table 1). All 20 prespecified comparisons pass Holm adjustment ($p \leq 0.006$) within the fixed 20-test comparison family under the stated sign-exchangeability assumption. Missing the target and emitting a wrong certificate remain distinct events.

[Development diagnostics.] A directed-GNN/MiniLM scorer audit averages three seeds within each of 296 graphs: wrong-certificate rates are at most 0.016 for CARTER versus 0.345–0.499 for pooling (Appendix B.2). Separately, a four-source TF–IDF sweep uses one seed at corruption 0.20. At $\alpha_s = 0.02$, both have zero observed wrong certificates and pooling certifies more edits; at the other tested levels, CARTER has a lower wrong-certificate rate (Appendix B.5). The sweep characterizes these fixed development settings.

3.3 RQ2: What does structural safety cost in automation?

[Sealed holdout.] CARTER provides additional partial repairs relative to the conservative full-candidate baselines. Its source-macro gain over Bonferroni and LTT is 0.35–0.79 edits per graph; automation of 0.405–0.754 means that at least one edit is certified, not that this proportion of all proposed edits is accepted (Table 1). Both baselines have zero observed wrong certificates.

Table 1: Sealed six-source holdout (224 test taxonomies, $\alpha_s = 0.10$). All entries are source-macro averages: five seeds within each graph, graphs within each source, and then six equally weighted source means. Coverage is whole-tree; "wrong" is the graph-level wrong-certificate rate; "edits" counts certified edits per graph, including deletions outside C. The planned coverage and wrong-rate contrasts against edge pooling have Holm-adjusted $p \leq 0.006$ at every rate.

	Coverage				Wrong-certificate rate			
Corruption rate	0.05	0.10	0.20	0.30	0.05	0.10	0.20	0.30
CARTER	0.929	0.935	0.928	0.915	0.001	0.000	0.000	0.001
Edge-pooled conformal	0.390	0.412	0.346	0.349	0.372	0.440	0.382	0.224
Bonferroni per-child	1.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000
Finite-population LTT	1.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000
	Certified edits per graph				Automation rate			
CARTER	0.596	1.066	1.554	2.223	0.405	0.628	0.657	0.754
Edge-pooled conformal	3.155	5.006	6.554	6.786	0.723	0.861	0.881	0.932
Bonferroni per-child	0.247	0.574	0.955	1.431	0.237	0.471	0.579	0.705
Finite-population LTT	0.247	0.574	0.955	1.431	0.237	0.471	0.579	0.705

Bonferroni's threshold and LTT's testing/fallback rule retain all candidates here (Appendix B). Their edits remove observed edges into the root, which lie outside C. These baselines establish the output of the full-candidate structural backbone in this evaluation.

Pooling produces more edits, but with a 0.224–0.440 source-macro graph-level wrong-certificate rate (Appendix Figure 2). Separately, minimum-edit point decoding improves edge F1 over no repair by 0.011–0.108; this is a reconstruction score, not a certification guarantee. Edit-count gains above are descriptive, without additional significance claims. Appendix G separately reports correct-edit counts and sensitivity to the two-graph ANZSRC source, preserving the six-source primary estimator.

3.4 RQ3: DOES A CALIBRATED FEEDBACK BUDGET PRESERVE LIFETIME SAFETY UNDER WRONG ANSWERS?

[Development-only programmed simulation.] A label-free hash split separates 148 calibration sessions from 148 evaluation sessions; the pipeline was frozen before generation (Appendix B.4). Independent-panel calibration selects $\hat{b}_{\mathrm{cert}} = 1$; persistent-child calibration separately selects 2.

The calibrated independent-panel arm retains the target at every prefix in 0.932 of sessions and emits a lifetime wrong edit in 2/148 sessions. It adds 2.095 certified edits per session over no feedback (two-sided 95% paired, dataset-stratified bootstrap $[1.872, 2.311]$; 5,000 resamples of graph sessions). Increments are positive in 127/148 sessions (Table 2). Neither failing session has its first wrong edit at the endpoint, illustrating why lifetime and terminal errors differ.

Matched-interface CRC gives the same states and certificates. Binary CRC on fixed session-exceedance scores selects exactly the same budget order statistic (Appendix B.6). The output-loss monotone envelope selects the same budgets of one/two on the saved calibration data. Its derived row reuses the same-budget outputs. Thus the certificate interface composes with either valid calibration route; Appendix B.6 also records why the historical raw output-loss adapter fails CRC's pointwise monotonicity condition.

The controls expose a safety–utility tradeoff on these same transcripts. No feedback has a retention rate of 0.953 with zero observed wrong edits; hard filtering has a retention rate of 0.736 and an error rate of 0.081, versus plurality's 0.088 error rate. Dawid–Skene matches hard filtering (Appendix B.8). Fixed $b_{\mathrm{cert}} = 2$ also has zero observed errors but adds only 0.216 edits. The calibrated output increment is 89% of the clean-answer increment ($2.095/2.351$). A post-hoc summary of saved outputs gives a canonical-correct edit increment of 2.088 (Appendix G); the controls here describe the output–safety tradeoff on fixed transcripts.

Table 2: Development-only simulated-answerer study ($n = 148$ evaluation sessions), with all arms replaying the same frozen transcripts. "Wrong" is any lifetime wrong certified edit, "retained" is target retention at every prefix, "conflict" means an empty budgeted set at some prefix, and "incremental" is the paired final-edit increase over no feedback. Edits include incorrect ones. The derived CRC-envelope row reuses the verified same-budget outputs; binary session-exceedance CRC also selects the same budget (Appendix B.6). Historical adapter records are reported there and in Appendix B.7.

Method	Source	Wrong	Retained	Edits	Conflict	Incremental
No feedback	—	0.000	0.953	1.345	0.000	0.000
Hard $b_{\mathrm{cert}} = 0$	—	0.081	0.736	2.872	0.236	1.527
Running plurality	—	0.088	0.851	3.486	0.007	2.142
Dawid–Skene + hard	Adapt. JRSS-C'79	0.081	0.736	2.872	0.236	1.527
CRC envelope (derived)	CRC rule	0.014	0.932	3.439	0.027	2.095
CARTER, calibrated $\hat{b}_{\mathrm{cert}} = 1$	ours	0.014	0.932	3.439	0.027	2.095
Fixed $b_{\mathrm{cert}} = 2$ sensitivity	sensitivity	0.000	0.953	1.561	0.007	0.216
Same-query clean answers	oracle	0.000	0.953	3.696	0.007	2.351

3.5 COMPUTATIONAL AND CALIBRATION DIAGNOSTICS

Exactness and cost. *[Development-only diagnostics.]* E1 matches coefficients, query values, and tie-breaks on 36 paired inputs; E3 matches objectives, edge classifications, and canonical witnesses on 12 reference-completed inputs. E1's median paired time ratio is 2.407 (reference/shared compiler), measuring computation and serialization under ambient load, not end-to-end latency. E3 includes witness solves; six 500-node jobs and two 100-node reference jobs exceed the 45-second whole-job limit (Appendix E).

Closed-loop query integration. E2 uses five-node graphs with target inclusion imposed, and 59 calibration/60 evaluation draws per policy and regime. Berlekamp yields 1.000/0.883 correct final edits per graph (iid/persistent), retaining the target at every prefix in 56/60 and 50/60 sessions, with 1/60 and 5/60 lifetime errors. Comparing its budget of one with entropy's two contrasts separately calibrated pipelines, not matched-risk policies. The post-hoc fixed-budget diagnostic reverses their utility ordering at budgets one through three (Appendix E); CRC-state matches each calibrated arm under shared budgets and transcripts.

Repeated calibration. To examine variation beyond E2's single fit, a separate development simulation repeats five-node calibration and evaluation 100 times. Every-prefix retention spans 0.971–0.979 across settings; lifetime error spans 0.006–0.010. Silent target loss, without a wrong emitted edit, spans 0.011–0.020. Persistent-error Berlekamp selects budgets one/two/three in 3/79/18 fits and produces 0.076 correct edits per graph, versus 0.498 under iid errors. The 100 outer fits, not pooled sessions or prefixes, are the independent resampling units; Appendix F gives paired outer-bootstrap intervals and all four adjusted pipeline contrasts, which show no adjusted advantage over entropy. These repetitions neither replace E2 nor guarantee nominal coverage for every calibration fit.

4 RELATED WORK

From parent prediction to repair. TaxoExpan learns attachments for new concepts while preserving existing relations (Shen et al., 2020). TaxoLLaMA fine-tunes a language model on WordNet for hypernym prediction (Moskvoretskii et al., 2024). Global decoding turns pairwise scores into a maximum-weight arborescence (Chen et al., 2021). These methods improve candidate generation or tree selection. CARTER complements this upstream role by certifying changes to an observed graph while retaining unresolved alternatives.

From coverage to retained states. Structured conformal prediction constructs compact prediction sets with coverage guarantees (Zhang et al., 2025). Interactive imitation learning handles expert shift and intermittent expert labels (Zhao et al., 2025); noise-robust online calibration corrects for known-rate uniform label noise (Xi et al., 2025). Their guarantees concern prediction-set coverage or

time-averaged errors. CRC controls expected bounded monotone loss (Angelopoulos et al., 2024); its session-budget special case connects to structured repair (Section 5).

From noisy search to exact certificates. Bayesian active learning gives query-cost guarantees under a probabilistic observation model (Golovin et al., 2010), whereas bounded-lie search specifies admissible answer errors (Pelc, 2002). For computation, tree constraints support generalized arc-consistency filtering (Fages & Lorca, 2011) and optimum branchings provide exact weighted-tree optimization (Edmonds, 1967). Exact force/forbid inference supplies family-wide edge certificates and opposing witnesses. Appendix Table 4 compares these output contracts.

5 DISCUSSION

Target retention makes exact certificates sound; correct edits need not imply retention, as programmed silent-loss diagnostics show (Section 3.5). Figure 1's schematic separates removing a wrong parent from choosing its replacement. Witnesses describe feasible alternatives, not target membership.

The central computational object is the cost-bounded tree family. Optimizing one tree supplies a witness, but certifying an edge requires checking the opposite structural restriction. Shared query compilation performs a different task: counting the family by disagreement layer. CARTER exposes these complementary operations through one interface, with the session budget selectable by conformal calibration or its matched binary-CRC formulation.

With the initial family and transcript fixed, increasing b_{cert} cannot add unanimous certificates between nonempty families. E2's post-hoc fixed-budget reversal cautions against policy-only interpretations of separately calibrated pipeline rankings. Repeated fits assess marginal, not per-fit, protection over calibration and evaluation draws; the event covers every prefix. Pipeline changes require rechecking assumptions, calibration, retention, wrong certificates, and correct edits.

6 LIMITATIONS

Task and evaluation. Certificates concern a frozen single-parent projection; deleted relations may remain valid in a source polyhierarchy. Fixed nodes and root exclude root repair, insertion, splitting, and merging. Structural tests simulate corruption of released taxonomies; feedback uses programmed answerers, not humans or LLMs. Five-node query and repeated-calibration studies impose initial inclusion. The six-source mean includes two ANZSRC test graphs (Appendix G); real-answerer performance and unseen-domain transfer require independent data.

Guarantee conditions. Every-prefix protection is marginal over calibration/evaluation draws, not conditional per fit. It requires a correct root, target-containing candidates, and exchangeable sessions from a frozen pipeline; budgets cannot recover missing edges. CARTER abstains on all six genuine MeSH/OEWN transitions; pooling's seed-averaged error is 0.667 (descriptive; Appendix B.11). Nominal recall 0.95 and realized child recall 0.942 yield a 0.419 session wrong-certificate rate (Appendix B.9). Candidate or pipeline changes require reassessment.

Computation. Certification and optional query counting have different costs. Counting is tested through 48 nodes; its arithmetic bound excludes integer bit growth. All six 500-node E3 jobs exceed the 45-second whole-job limit (Appendix E). The 152-second reference observation and 3.58-second optimized median use unmatched inputs and timing units (Appendix B.12). Release-scale interactive latency remains unestablished, especially for dense candidate sets.

7 CONCLUSION

CARTER connects calibrated primary-parent tree families to complete edge certificates and opposing witnesses. With a correct root, target-containing candidates, and exchangeable sessions from a frozen pipeline, its marginal guarantee protects target retention and all certified edits at every prefix. Whole-tree calibration improves coverage and reduces wrong certificates relative to edge pooling on the sealed structural benchmark. Separate programmed-feedback and repeated-calibration studies show why target retention and correct edits need joint assessment. These results support auditable partial repair, with feasible alternatives explaining unresolved decisions.

ETHICS STATEMENT

The benchmark experiments use released taxonomy data under their stated licenses. The interaction experiments use declared simulated error regimes rather than human or AI-generated answers as expert ground truth. Earlier expert-annotation pilot claims are excluded from the evidence supporting this paper because the available records do not establish their human-annotator provenance (Appendix B.1). The anonymous supplement excludes individual annotation responses and participant-identifying materials.

REPRODUCIBILITY STATEMENT

The appendix specifies assumptions, proofs, generators, and evidence scope. The anonymous supplement provides hashes, configurations, saved decisions, witnesses, tests, and study-specific environments. `REPRODUCE.md` gives separate commands for payload verification, CPU smoke tests, released-result reaggregation, and fresh development replay, with non-overwriting outputs. Frozen per-graph tables support recomputation of the original source-macro intervals and paired tests. The sealed integrity check verifies downstream certificates from frozen score transcripts; it does not reopen the consumed holdout or independently rerun or retrain the historical scorer. Historical sealed-run pins and the repeated-calibration environment are documented separately; Appendix C records invalidated and superseded attempts. The numeric audit identifies source matching, reaggregation, and actual reexecution as distinct verification levels. Source matching alone is not fresh experimental replication.

AI USE STATEMENT

Generative AI tools supported literature search; conceptual and theorem review, formulation of mathematical claims, and proof drafting; experiment design, method implementation and testing, including synthetic-graph and programmed-feedback generators; interpretation of computed results; and manuscript drafting, revision, translation, schematic figure editing, and data-artifact preparation. The authors independently verified all theorems and proofs line by line. Reported experimental values are outputs of executed code with saved artifacts, not model-generated prose. Synthetic graphs and programmed feedback were produced by executed generators; no LLM-authored answers were attributed to people, and no generative model served as an experimental expert or ground truth. Implementation checks include reference-solver comparison and small-instance enumeration. The authors retain responsibility for the final text, proofs, claims, citations, code, and data.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, volume 2024, pp. 55198–55218, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Anastasios N. Angelopoulos. Conformal risk control for non-monotonic losses, 2026. URL https://arxiv.org/abs/2602.20151v1.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Australian Bureau of Statistics. Australian and new zealand standard research classification (ANZSRC), 2020: Fields of research, 2020. URL https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-research-classification-anzsrc/latest-release. Corrected release dated 2025-10-24; source file ANZSRC2020_for.xlsx; SHA-256 92B94664EB1E43DB1CBCAEB54E60575E3F8573CD10E5E4BB3A9C806CBE462E35; CC BY 4.0.

Walter Baur and Volker Strassen. The complexity of partial derivatives. *Theoretical Computer Science*, 22(3):317–330, 1983. doi: 10.1016/0304-3975(83)90110-X. URL https://doi.org/10.1016/0304-3975(83)90110-X.

Stuart J. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Information Processing Letters*, 18(3):147–150, 1984. doi: 10.1016/0020-0190(84)90018-8. URL https://doi.org/10.1016/0020-0190(84)90018-8.

Catherine Chen, Kevin Lin, and Dan Klein. Constructing taxonomies from pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4687–4700. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.373. URL https://aclanthology.org/2021.naacl-main.373/.

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

Alexander Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 28 (1):20–28, 1979. doi: 10.2307/2346806. URL https://doi.org/10.2307/2346806.

Jack Edmonds. Optimum branchings. *Journal of Research of the National Bureau of Standards, Section B*, 71B(4):233–240, 1967. doi: 10.6028/jres.071B.032. URL https://nvlpubs.nist.gov/nistpubs/jres/71b/jresv71bn4p233_a1b.pdf.

Jean-Guillaume Fages and Xavier Lorca. Revisiting the tree constraint. In *Principles and Practice of Constraint Programming – CP 2011*, volume 6876 of *Lecture Notes in Computer Science*, pp. 271–285. Springer, 2011. doi: 10.1007/978-3-642-23786-7_22. URL https://doi.org/10.1007/978-3-642-23786-7_22.

Food and Agriculture Organization of the United Nations. AGROVOC multilingual thesaurus, 2026. URL https://www.fao.org/agrovoc/access. Project snapshot file agrovoc_core.rdf.zip; SHA-256 561A0E18E4A5BEE3CD41B01316442D4BA5353A89D5AD7245438583D8B6B93734; CC BY 3.0 IGO; accessed 2026-07-31.

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

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Terry Koo, Amir Globerson, Xavier Carreras, and Michael Collins. Structured prediction models via the matrix-tree theorem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 141–150, Prague, Czech Republic, 2007. Association for Computational Linguistics. URL https://aclanthology.org/D07-1015/.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.2017.1307116.

John P. McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets*, pp. 14–19. The European Language Resources Association (ELRA), 2020. URL https://aclanthology.org/2020.mmw-1.3/.

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. TaxoLLaMA: WordNet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2331–2350, 2024. doi: 10.18653/v1/2024.acl-long.127. URL https://aclanthology.org/2024.acl-long.127/. ACL Anthology 2024.acl-long.127.

Open English WordNet Community. Open English WordNet 2024. https://en-word.net/downloads, 2024. File english-wordnet-2024.xml.gz; SHA-256 E1F633B0A93758CAE34EA27C44C4DAD310A8AF2467B155F99DD6673AF697E875; CC BY 4.0.

Open English WordNet Community. Open English WordNet 2025. https://en-word.net/downloads, 2025. Release dated 2025-12-31; file english-wordnet-2025.xml.gz; SHA-256 9CA6D1DCB75F822FDD66617F7D9DA48142ACE38DD544D6AD5E2FECA1674AD3FE; CC BY 4.0.

Andrzej Pelc. Searching games with errors—fifty years of coping with liars. *Theoretical Computer Science*, 270(1–2):71–109, 2002. doi: 10.1016/S0304-3975(01)00303-6. URL https://doi.org/10.1016/S0304-3975(01)00303-6.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html.

Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. TaxoExpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *Proceedings of The Web Conference 2020*, pp. 486–497. Association for Computing Machinery, 2020. doi: 10.1145/3366423.3380132. URL https://doi.org/10.1145/3366423.3380132.

UNESCO. UNESCO thesaurus, 2026. URL https://vocabularies.unesco.org/unesco/en/. Project snapshot last modified 2026-06-29; file unesco-thesaurus.ttl; SHA-256 698B47AE257EF4D15BD3A4C101C451C2DE9F20FC497BECEC9E665623BD05450D; CC BY-SA 3.0 IGO.

U.S. National Library of Medicine. Medical subject headings (MeSH), 2025 production year, 2025. URL https://nlmpubs.nlm.nih.gov/projects/mesh/2025/xmlmesh/desc2025.gz. File mesh-desc2025.gz; SHA-256 99285A75544779DE23747F062830B8E36746BDFC5FC44A25ECE5855172439C8F; reuse governed by NLM terms and conditions.

U.S. National Library of Medicine. Medical subject headings (MeSH), 2026 production year, 2026. URL https://www.nlm.nih.gov/databases/download/mesh.html. File mesh-desc2026.gz; SHA-256 9FE35B3170652376A592DAF69E91A80D6C693ECAF9C571CEB701D04204CB357D; reuse governed by NLM terms and conditions.

Gilles Villard. Kaltofen's division-free determinant algorithm differentiated for matrix adjoint computation. *Journal of Symbolic Computation*, 46(7):773–790, 2011. doi: 10.1016/j.jsc.2010.08.012. URL https://doi.org/10.1016/j.jsc.2010.08.012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer New York, New York, NY, 1 edition, 2005. doi: 10.1007/b106715. URL https://link.springer.com/book/10.1007/b106715.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Huajun Xi, Kangdao Liu, Hao Zeng, Wenguang Sun, and Hongxin Wei. Exploring the noise robustness of online conformal prediction. In *Advances in Neural Information Processing Systems*,

volume 38, 2025. doi: 10.52202/085713-1546. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/423d0909791493b7c10916fd328c2913-Abstract-Conference.html. arXiv:2501.18363.

Botong Zhang, Shuo Li, and Osbert Bastani. Conformal structured prediction. In *International Conference on Learning Representations*, volume 2025, pp. 8563–8583, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1868a3c73d0d2a44c42458575fa8514c-Abstract-Conference.html.

Michelle Zhao, Henny Admoni, Reid Simmons, Aaditya Ramdas, and Andrea Bajcsy. Conformalized interactive imitation learning: Handling expert shift and intermittent feedback. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/19deda616aebc9bee6c1990baa5f1f0e-Abstract-Conference.html.

Ran Zmigrod, Tim Vieira, and Ryan Cotterell. Efficient computation of expectations under spanning tree distributions. *Transactions of the Association for Computational Linguistics*, 9:675–690, 2021. doi: 10.1162/tacl_a_00391. URL https://aclanthology.org/2021.tacl-1.41/.

A Assumptions, proofs, and exactness checks

A.1 Notation

Table 3: Symbols used in the main text. Numerical settings are study-specific; the parameter registry binds them to saved configurations and development outputs.

Symbol	Meaning
$V,\ r$	Concepts of one taxonomy and its designated root
C	Frozen candidate-edge universe on V
$\mathcal{T}(G, r)$	Arborescences of G rooted at r
$T^\star$	Canonical primary-parent target tree
$s(e)$	Frozen edge nonconformity score
$S(T)$	Complete-tree bottleneck score $\max_{e\in E(T)} s(e)$
q	Raw split-conformal threshold
q'	Operational threshold after graph-only feasibility closure
$D_0,\ H_0$	Threshold-allowed edges and the allowed graph (V, D_0)
$\mathcal{V}_0$	Initial version space $\mathcal{T}(H_0, r)$
α_s	Structural miscoverage level; 0.10 sealed, 0.05 paired study
$v_s,\ A_s$	Child queried at round s and the returned acceptable-parent set
$\tau,\ \tau_j$	Paid questions in the next session and calibration session j
$H_{\max}$	Frozen maximum session horizon (4 in the canonical feedback study)
$L_t(T)$	Answers among the first t that T contradicts
b_{cert}	Calibrated certificate budget, Equation (4)
$Y_j,\ Y_{(k)}$	Complete-session exclusion count and its k-th calibration order statistic
δ_h	Session-budget miscoverage level (0.05)
$\mathcal{V}_t^{b_{\text{cert}}}$	Budgeted version space at round t
$M_t,\ P_t,\ I_t$	Mandatory, possible, and impossible candidate edges
$E(T),\ E_{\text{obs}}$	Tree edges and observed taxonomy edges before certified repair
$c_t(e)$	Answers ruling out edge e, Equation (6)
b_{qry}	Frozen query-utility budget; 1 in the independent utility audit, $\{0, 1, 2, 3\}$ in scaling; absent from the canonical ambiguity-block policy
$n_i,\ m_i(e)$	Error-layer counts and their per-edge refinements
$\mathcal{S}_i(\rho),\ L(T; \rho)$	Error layer and contradiction count for binary transcript ρ
$W_h(\rho)$	Berlekamp transcript weight with h questions remaining
$B_\ell(h)$	Radius-ℓ binary-ball volume; zero for $\ell < 0$
$c_\rho,\ Z_\rho,\ M_{\rho,e}$	Transcript edge costs, tree-count polynomial, and edge-layer polynomial
$F(\mathbf{x}),\ x_e,\ d$	Tree-generating polynomial, edge variable, and tree size $d = \lvert V\rvert - 1$

A.2 PROBABILITY SPACE AND GRAPH-INDUCED INITIAL SET

Freeze the edge scorer and candidate-universe rule on training data disjoint from both calibration streams. Structural calibration uses labeled calibration targets to determine the raw threshold q. The subsequent graph-only feasibility closure $q' \geq q$ uses no target labels of the new instance. Conditional on that first stream and the resulting frozen procedure, reliability-calibration sessions and the next deployment session are assumed exchangeable. Before collecting them, fix the measurable query policy (including its randomization and any utility budget b_{qry}), a deterministic maximum horizon $H_{\max} \in \mathbb{Z}_{\geq 0}$, a stopping time $\tau \leq H_{\max}$ based only on information available so far (the filtration generated by past feedback and internal randomization), and the answer-aggregation protocol. A complete taxonomy–answerer interaction produced by this pipeline is one session; answers within a session need not be independent.

Let

$$C \subseteq \{(u, v) \in V \times V : u \neq v,\ v \neq r\}$$

be the frozen simple directed candidate-edge universe, and assume that this universe contains the canonical target: $T^\star \in \mathcal{T}((V, C), r)$ almost surely. Also assume $0 < \alpha_s, \delta_h$ with $\alpha_s + \delta_h < 1$. If the candidate universe does not contain the target, the target-inclusion assumption is violated and the theorem is out of scope; conformal calibration cannot repair this retrieval failure. The score is bottleneck-factorized:

$$S(T) = \max_{e \in E(T)} s(e).$$

Let $R_1, \ldots, R_{m_s}, R_{m_s+1}$ be exchangeable whole-taxonomy calibration/deployment scores, where $R_j = S_j(T_j^\star)$, and set

$$k_s = \lceil (m_s + 1)(1 - \alpha_s) \rceil .$$

If $k_s \leq m_s$, let q be the k_s-th calibration order statistic; if $k_s = m_s + 1$, set $q = +\infty$, so all edges in C are allowed. For the proof, attach iid continuous tie-breakers to all $m_s + 1$ scores. When $k_s \leq m_s$, the strict event $R_{m_s+1} > q$ implies a randomized rank greater than k_s, and hence

$$\Pr\{R_{m_s+1} > q\} \leq \frac{m_s + 1 - k_s}{m_s + 1} \leq \alpha_s.$$

When $k_s = m_s + 1$, including the target in the candidate universe makes coverage deterministic. Removing the proof-only tie-breakers is conservative for the strict failure event.

A feasibility closure $q' \geq q$ may be computed from the observed, unlabeled instance only. Define

$$D_0 = \{e \in C : s(e) \leq q'\}, \qquad H_0 = (V, D_0), \qquad \mathcal{V}_0 = \mathcal{T}(H_0, r).$$

Because feasibility closure only enlarges the set obtained at q, it cannot reduce coverage; therefore $\Pr\{T^\star \notin \mathcal{V}_0\} \leq \alpha_s$ over repeated draws of the structural calibration sample and the next taxonomy. The efficient certificate and query compiler below require this graph-induced form. For an arbitrary covered structured set, the lifetime containment proof still applies, but force/forbid optimization and counting must be restricted to that set; an ordinary arborescence solver or Matrix–Tree determinant is not sufficient.

The structural calibration stream is completed first. Its realized threshold rule and query policy are frozen before a disjoint reliability calibration stream is collected. The lifetime probability below is marginal over both calibration samples and the next session; its union bound does not require the two failure events to be independent.

Throughout, $T^\star$ is one predeclared canonical primary-parent tree. In the released benchmark construction, edges internal to nontrivial strongly connected components are first removed to obtain an acyclic hierarchy. All remaining source roots are then attached to a dataset-scoped virtual root. For each non-root node, the canonical parent is then the source parent with minimum depth from that virtual root, with deterministic node-identifier order breaking remaining ties. This projection defines the evaluation target for the single-parent certificate. Correctness of an edit means agreement with $T^\star$. Additional acceptable parents can improve efficiency and support a separate semantic metric; absence from $T^\star$ asserts only exclusion from the canonical tree, not semantic unacceptability.

A.3 SESSION CALIBRATION AND LIFETIME CONTAINMENT

Whole-session order statistic. Let $Y_1, \dots, Y_n, Y_{n+1} \in \{0, \dots, H_{\max}\}$ be exchangeable complete-session target-parent exclusion counts $Y_j = L_{\tau_j}(T_j^\star)$ from the frozen pipeline, and set

$$k = \lceil (n+1)(1-\delta_h) \rceil .$$

If $k = n+1$, set $b_{\text{cert}} = H_{\max}$; the failure event is then impossible. Otherwise let $b_{\text{cert}} = Y_{(k)}$, the k-th smallest calibration count. For the proof, attach iid continuous tie-breakers to all $n+1$ counts. Exchangeability makes the randomized rank of the next count uniform on $\{1, \dots, n+1\}$. The strict event $Y_{n+1} > b_{\text{cert}}$ implies a rank greater than k, hence

$$\Pr\{Y_{n+1} > b_{\text{cert}}\} \le \frac{n+1-k}{n+1} \le \delta_h.$$

Removing the proof-only tie randomization can only make the upper prediction bound more conservative. This probability is marginal over both the calibration sample and the next session; it is not conditional on the realized b_{cert}. The separately frozen b_{qry} may affect which questions generate these counts, but it is not recalibrated from this stream and need not equal b_{cert}.

Lifetime theorem. Let $E_s = \{T^\star \in \mathcal{V}_0\}$ and $E_h = \{L_\tau(T^\star) \le b_{\text{cert}}\}$. A union bound, with no independence assumption between these events, gives

$$\Pr(E_s \cap E_h) \ge 1 - \alpha_s - \delta_h.$$

On this event, nonnegativity of each target-parent exclusion loss gives $L_t(T^\star) \le L_\tau(T^\star) \le b_{\text{cert}}$ for every paid-query prefix $t \le \tau$, so $T^\star \in \mathcal{V}_t^{b_{\text{cert}}}$ simultaneously for all such t. Because $T^\star$ is itself one of the retained trees, any edge present in every retained tree is present in $T^\star$, and any edge present in none is absent from $T^\star$; certified additions and deletions are therefore correct. Every round and edge decision follows from this same event, so no round-wise or edge-wise multiplicity correction is needed.

Frozen-transcript audit. The finite-population alternative first freezes all N answers, then draws a simple random sample of n positions without replacement. It assumes the independent adjudication is correct and that no sampled answer is selectively discarded. For a fixed transcript with M target-parent exclusions, the audited count has a $\text{Hypergeom}(N, M, n)$ distribution. If k sampled answers are wrong, set

$$U_\delta(k) = \max\{M \in \{0, 1, \dots, N\} : \Pr[\text{Hypergeom}(N, M, n) \le k] > \delta\}, \tag{7}$$

for $0 < \delta < 1$. Inverting the lower tail with this strict inequality gives a finite-sample upper confidence bound satisfying $\Pr_M\{M \le U_\delta(K)\} \ge 1 - \delta$, where $K \sim \text{Hypergeom}(N, M, n)$. Because the sampling distribution is discrete, its coverage can exceed $1 - \delta$. Consequently, at most $U_\delta(k) - k$ errors remain after the k audited errors are corrected. The bound applies only to a final recomputed certificate; it does not retroactively validate online displays.

A.4 COMPLETE FORCE/FORBID CERTIFICATE

At each paid-query prefix, a nonabstaining response is a nonempty acceptable parent set. An abstention is encoded as the full allowed-parent set $\{u : (u, v_s) \in D_0\}$; it therefore adds zero loss to every tree in $\mathcal{V}_0$ while still consuming one round of the bounded horizon.

For $e = (u, v) \in D_0$, define

$$c_t(e) = \sum_{s \le t \,:\, v_s = v} \mathbf{1}\{u \notin A_s\}.$$

Each rooted arborescence selects exactly one incoming edge for every non-root child. Interchanging the two finite sums therefore gives

$$\sum_{e \in E(T)} c_t(e) = \sum_{s=1}^{t} \mathbf{1}\{\text{pa}_T(v_s) \notin A_s\} = L_t(T).$$

Thus $\mathcal{V}_t^{b_{\text{cert}}}$ is exactly the set of cost-$\le b_{\text{cert}}$ rooted arborescences of H_0.

Let T_0 minimize this cost. If its cost exceeds b_{cert}, no member of $\mathcal{V}_t^{b_{\text{cert}}}$ exists, so the implementation fails closed. Otherwise:

- If $e \in T_0$, a budget-feasible tree excluding e exists if and only if the forbid-e problem is feasible and its optimum is at most b_{cert}. The absence of such a tree makes e mandatory.
- If $e \notin T_0$, a budget-feasible tree including e exists if and only if the force-e problem is feasible and its optimum is at most b_{cert}. The absence of such a tree makes e impossible.

An infeasible restricted problem is treated as having an optimum of $+\infty$, never as a solver failure. In every ambiguous case, T_0 supplies one side of the decision and the restricted optimum supplies the other. One baseline solve and one opposite restriction per allowed edge use at most $|D_0| + 1$ oracle calls. Edges in $C \setminus D_0$ are absent from every tree in $\mathcal{V}_0$ and are impossible without an additional solve. For edit reporting, observed edges outside C are also excluded from every candidate tree. Certified deletions are therefore $E_{\mathrm{obs}} \setminus P_t = (E_{\mathrm{obs}} \cap I_t) \cup (E_{\mathrm{obs}} \setminus C)$. These exclusions require no additional solve; their correctness follows from the existing assumption that the target lies in C.

The possible-edge union is not itself the robust version space: combining individually possible edges can violate the tree or budget constraint. A point recommendation must therefore belong to $\mathcal{V}_0$ and have total disagreement at most b_{cert}. Optimizing only over the union graph P_t does not enforce these conditions and remains uncertified.

A.5 Binary-to-set bridge and branch recurrence

The binary query layer uses the separately frozen utility budget $b_{\mathrm{qry}} \in \mathbb{Z}_{\geq 0}$, not the calibrated certificate budget $b_{\mathrm{cert}} \in \mathbb{Z}_{\geq 0}$. It compiles nonabstaining binary answers. A paid abstention leaves the transcript state unchanged and decrements the remaining horizon; the stated minimax rule covers the two answer branches, and adversarial abstention sits outside its scope. Define the operational query family

$$\mathcal{Q}_0 = \{(u, v) \in D_0 : |\{p : (p, v) \in D_0\}| \geq 2\}.$$

For $e = (u, v) \in \mathcal{Q}_0$, the binary query is $q_e(T) = \mathbf{1}\{e \in T\}$. Map answer $a = 1$ to $A = \{u\}$, and answer $a = 0$ to

$$A = \{p : (p, v) \in D_0,\ p \neq u\}.$$

The definition of $\mathcal{Q}_0$ makes both answer sets nonempty. Because a tree chooses exactly one parent of v, its set-valued exclusion loss equals $\mathbf{1}\{q_e(T) \neq a\}$. This proves that the robust-certificate and liar-compiler costs are the same object under these binary questions.

Write ρ for a binary transcript, distinct from the stopping time τ, consisting of a finite sequence of answered queries $(e_1, a_1), \ldots, (e_t, a_t)$ with $e_j \in \mathcal{Q}_0$, and let $L(T; \rho) = \sum_{j \leq t} \mathbf{1}\{q_{e_j}(T) \neq a_j\}$ be the number of those answers that T contradicts. Define the *error layers*

$$\mathcal{S}_i(\rho) = \{T \in \mathcal{V}_0 : L(T; \rho) = i\}, \qquad i = 0, 1, \ldots,$$

so $\mathcal{S}_i(\rho)$ collects the trees that contradict exactly i answers. For any integer $b \geq 0$, $\mathcal{V}_t^b = \bigcup_{i=0}^{b} \mathcal{S}_i(\rho)$. For a remaining horizon h, the ***Berlekamp weight*** of the transcript is

$$W_h(\rho) = \sum_{i \geq 0} |\mathcal{S}_i(\rho)|\, B_{b_{\mathrm{qry}} - i}(h),$$

the number of feasible tree–answer-continuation pairs, counted with multiplicity across trees. It is not the number of distinct answer strings: one continuation can remain compatible with several trees. $B_\ell(h)$ is defined below. Let

$$n_i = |\mathcal{S}_i(\rho)|, \qquad m_i(e) = |\{T \in \mathcal{S}_i(\rho) : e \in T\}|,$$

with $n_{-1} = m_{-1} = 0$. After answer 1, layer i contains the previous layer-i trees containing e and previous layer-$i - 1$ trees excluding it:

$$n_i^{(1)} = m_i + n_{i-1} - m_{i-1}.$$

After answer 0, it contains the previous layer-i trees excluding e and previous layer-$i - 1$ trees containing it:

$$n_i^{(0)} = n_i - m_i + m_{i-1}.$$

Set $B_\ell(h) = \sum_{j=0}^{\ell} \binom{h}{j}$ for $\ell \geq 0$, using the standard convention $\binom{h}{j} = 0$ for $j > h$, and $B_\ell(h) = 0$ otherwise. Substituting the two recurrences and applying Pascal's identity gives

$$W_{h-1}(\rho \cdot (e, 1)) + W_{h-1}(\rho \cdot (e, 0)) = W_h(\rho).$$

Minimizing the larger child weight is therefore the exact one-step Berlekamp-weight minimax rule within the allowed single-edge query family. It does not imply full-horizon optimality. For an integer $h \geq 1$, this rule returns a query only when $\mathcal{Q}_0 \neq \varnothing$; otherwise the interface reports no admissible binary query and stops.

Theorem 3 (single-circuit exact query compilation). Fix an integer remaining horizon $h \geq 1$. Given the edge-cost vector c_ρ, all layer counts through b_{qry} and both horizon- $h-1$ answer weights for every admissible edge are computed by one division-free determinant circuit and one reverse accumulation, in $O(d^4(b_{\mathrm{qry}}+1)^2 + |D_0|(b_{\mathrm{qry}}+1))$ exact-integer operations, where $d = |V|-1$ is the number of non-root concepts. This bound covers the forward determinant circuit, its reverse accumulation, and materialization of all per-edge layer coefficients and answer weights after c_ρ has been supplied. Building c_ρ directly from a length-t transcript costs a further $O((t+1)|D_0|)$, so the end-to-end per-round cost is the sum of these two costs. The bound is pseudo-polynomial in b_{qry}, and the stored tape occupies $O(d^4(b_{\mathrm{qry}}+1))$ integer slots. Weighted tree marginals and division-free adjoint circuits are prior work (Koo et al., 2007; Baur & Strassen, 1983; Berkowitz, 1984; Villard, 2011; Zmigrod et al., 2021); the contribution is their exact-integer, bounded-lie specialization.

Layer-projection identity. Put $\bar{h} = h-1$,

$$s_e = \sum_{i=0}^{b_{\mathrm{qry}}} m_i(e)\binom{\bar{h}}{b_{\mathrm{qry}}-i}, \qquad U_{\mathrm{lay}} = \sum_{i=0}^{b_{\mathrm{qry}}} n_i\binom{\bar{h}}{b_{\mathrm{qry}}-i},$$

and

$$A_{\mathrm{com}} = \sum_{i=0}^{b_{\mathrm{qry}}} n_i B_{b_{\mathrm{qry}}-i-1}(\bar{h}).$$

Substituting the branch recurrences and using $B_k(\bar{h}) - B_{k-1}(\bar{h}) = \binom{\bar{h}}{k}$ gives

$$W_{\bar{h}}(\rho\cdot(e,1)) = A_{\mathrm{com}} + s_e, \qquad W_{\bar{h}}(\rho\cdot(e,0)) = A_{\mathrm{com}} + U_{\mathrm{lay}} - s_e.$$

Thus one-step minimax is equivalently $\arg\min_e |2s_e - U_{\mathrm{lay}}|$. At an empty transcript only layer zero is occupied, so s_e and U_{lay} are ordinary edge-inclusion and version counts multiplied by the common factor $\binom{\bar{h}}{b_{\mathrm{qry}}}$. If that factor is positive, uniform balance and Berlekamp balance therefore agree; if it is zero, all queries tie. This explains why an empty-prefix experiment cannot establish a strict advantage over uniform entropy. With a nonuniform layer state, the common-factor argument no longer applies.

A.6 Directed Matrix–Tree compiler

Writing the transcript as $\rho = ((e_s, a_s))_{s=1}^t$ with $e_s = (u_s, v_s)$, assign candidate edge $f = (x, y)$ the cost

$$c_\rho(f) = \sum_{s:a_s=0} \mathbf{1}\{f = e_s\} + \sum_{s:a_s=1} \mathbf{1}\{y = v_s,\ f \neq e_s\}, \tag{8}$$

so that $C_\rho(T) = L(T;\rho) = \sum_{f\in E(T)} c_\rho(f)$. Give allowed edge $u \to v$ weight $w_{uv} = z^{c_\rho(u,v)}$. Define the child-row, parent-column in-degree Laplacian

$$L_{vv} = \sum_{u:(u,v)\in D_0} w_{uv}, \qquad L_{vu} = -w_{uv} \quad (u \neq v),$$

with zero off-diagonal entries when $u \to v \notin D_0$. Delete the row and column indexed by root r, obtaining $Q_\rho(z)$. The directed Matrix–Tree identity for this orientation is

$$\det Q_\rho(z) = \sum_{T\in\mathcal{T}(H_0,r)} \prod_{e\in E(T)} w_e = \sum_{T\in\mathcal{T}(H_0,r)} z^{C_\rho(T)}.$$

The transcript-to-edge formula in Equation (8) charges a negative answer exactly when its queried edge is selected, and a positive answer exactly when an alternative incoming edge for the same child is selected. Hence the exponent equals $C_\rho(T)$, and the coefficient of z^i is n_i.

To count trees containing $e = (u, v)$, delete every other incoming edge of v before forming the same minor. Every resulting arborescence contains e, and every original arborescence containing e remains,

so the coefficient of z^i is exactly $m_i(e)$. If the restriction admits no tree, the determinant is the zero polynomial. This forced-determinant route is retained as an independent reference oracle.

For the single-circuit implementation, assign a distinct formal input x_e to every allowed edge and write

$$F(\mathbf{x}) = \det Q(\mathbf{x}) = \sum_{T} \prod_{f \in E(T)} x_f.$$

This fixed-graph polynomial is the F of the main text; transcript dependence enters through the input substitution below. It is multilinear in every edge input, hence

$$x_e \frac{\partial F}{\partial x_e} = \sum_{T : e \in E(T)} \prod_{f \in E(T)} x_f.$$

Evaluate the same division-free determinant circuit at $x_f = z^{c_\rho(f)}$ and run reverse accumulation through its additions, subtractions, and multiplications. The forward value is $Z_\rho(z)$, while the edge input's adjoint multiplied by x_e is $M_{\rho,e}(z)$. This obtains all edge-layer polynomials from one forward–reverse circuit instead of $|D_0|$ additional determinants.

All determinant operations take place in $R_{b_{\mathrm{qry}}} = \mathbb{Z}[z]/(z^{b_{\mathrm{qry}}+1})$. The quotient map from $\mathbb{Z}[z]$ to $R_{b_{\mathrm{qry}}}$ is a ring homomorphism; additions and multiplications cannot make a discarded term of degree greater than b_{qry} contribute to a coefficient of degree at most b_{qry}. Therefore computing the determinant and its formal reverse sweep in $R_{b_{\mathrm{qry}}}$ preserves all required layers. The reverse rules use no inverse or cancellation, and Berkowitz's division-free circuit is valid over any commutative ring, including this quotient ring with zero divisors (Baur & Strassen, 1983; Berkowitz, 1984; Villard, 2011).

Let $d = |V| - 1$. The reference Berkowitz circuit uses $O(d^4)$ ring operations per determinant. Naive truncated-polynomial addition and multiplication cost $O(b_{\mathrm{qry}} + 1)$ and $O((b_{\mathrm{qry}} + 1)^2)$ integer arithmetic operations, respectively. Recording that circuit once and reversing every operation has constant-factor ring-operation overhead; multiplying each edge adjoint by its monomial is a degree shift. The dominant per-round bound after cost compilation is

$$O\big(d^4(b_{\mathrm{qry}} + 1)^2 + |D_0|(b_{\mathrm{qry}} + 1)\big).$$

A naive tape retains $O(d^4(b_{\mathrm{qry}} + 1))$ integer coefficient slots. The forced-determinant reference instead has the older $O((|D_0| + 1)d^4(b_{\mathrm{qry}} + 1)^2)$ arithmetic-operation bound. A direct transcript compiler costs $O(|D_0| + t|D_0|) = O((t + 1)|D_0|)$ at round t. Without cross-round reuse, an $H_{\max}$-round session is bounded by the sum of these per-round costs. This is pseudo-polynomial in the numeric value of b_{qry}, not polynomial in its bit length, and not a bit-complexity or strong-polynomial result: exact coefficient bit lengths grow with the number of arborescences.

A.7 Necessary query bound and restricted-query counterexample

Each candidate target consistent with i past disagreements occupies $\sum_{j=0}^{b_{\mathrm{qry}}-i} \binom{h}{j}$ answer strings of length h. Disjoint decoding under b_{qry} lies therefore requires $W_h(\rho) \le 2^h$. Initially, for $M = |\mathcal{T}(H_0, r)|$, any ℓ-query strategy must satisfy

$$M \sum_{j=0}^{b_{\mathrm{qry}}} \binom{\ell}{j} \le 2^\ell.$$

This is necessary, not sufficient, because queries are restricted to parent edges. In the d-parent star construction, every allowed query splits the d targets as $1 : (d - 1)$, so the noiseless worst case needs $d - 1$ queries. No unconditional logarithmic query bound is claimed.

B Development validation and scope boundaries

The forced-determinant reference is compared with explicit tree enumeration for every candidate subgraph through four nodes and randomized graphs through seven nodes. The forward–reverse compiler is compared coefficient by coefficient with both that reference and enumeration on randomized

Table 4: Selected interface comparisons verified against the cited sources. The last column describes our instantiated object, not an impossibility claim about extending another method or an exhaustive novelty search.

Selected work	Returned object	Relation to the present interface
Structured conformal (Zhang et al., 2025)	Compact structured representations of prediction sets	We retain a graph-induced family of complete primary-parent trees and derive statements shared by that family.
Taxonomy decoding (Chen et al., 2021)	Pairwise parent scores reconciled into a maximum-weight arborescence	Candidate scoring is complementary; our output is a partial certificate with explicit unresolved alternatives, not one best tree.
Conformal risk control (Angelopoulos et al., 2024)	A parameter selected to control expected monotone bounded loss	Binary session-exceedance loss gives the same budget as our order statistic. Using the same transcript and certificate solver also gives the same states and witnesses; it is not an empirical superiority comparison.
Weighted-tree inference (Koo et al., 2007; Zmigrod et al., 2021)	Partition functions, edge marginals, and expectations	We specialize the established derivative identity to repeated-transcript integer error layers using a division-free circuit; the identity is not new.
CARTER	Implicit rooted-tree version space	Combines separately calibrated session exclusions, complete edge certificates, and compiled one-step query values

graphs through six nodes, including repeated contradictory queries and budgets zero through three. Its general polynomial determinant is independently compared with Leibniz expansion through size six. For every non-root child v, edge polynomials satisfy, coefficient by coefficient,

$$\sum_{e \to v} M_e(z) = Z(z), \qquad \sum_{e \in D_0} M_e(z) = (|V| - 1)Z(z).$$

An executable four-tree counterexample has state counts $(2, 2)$. Two queries both split the four surviving trees 2:2, but their projected signals are $s_e = 2$ and $s_e = 3$ with $A_{\mathrm{com}} = 2, U_{\mathrm{lay}} = 6$; uniform entropy selects the former by its frozen lexical tie-break and incurs worst child weight six, whereas the Berlekamp rule selects the latter with worst weight five. The development query bundle enumerates every legal answer path with at most one lie for 708 truth-tree blocks on 16 five-node graphs; every path retains its target. After three paid queries, the Berlekamp rule reduces the mean worst-case terminal version-space count from 46.455 to 23.297, compared with 35.194 for the frozen-random policy. It ties uniform version-space entropy at 23.297 in this grid; the constructed nonuniform-layer case above, rather than this aggregate, isolates their difference. Every child has three allowed parents, so the evaluated query family satisfies $\mathcal{Q}_0 = D_0$; a separate regression test excludes singleton-parent edges from the set-valued bridge. Its artifact-set SHA-256 is `b23d402ca8a4a6f2b43f25bda3c46a8cd80a1284ccddce3a334f7f25d70686b6`. These are query–transcript–version-space implementation and development checks, not certificate measurements or evidence from the sealed confirmatory holdout.

Canonical feedback run. The canonical Table 2 and RQ3 results come from `joint_e2e_public_dev_v1`, reported at `scope=overall`, `regime=iid_panel`, `policy=ambiguity_block`. This policy ranks unresolved children by the number of allowed parents; it is not the binary Berlekamp compiler studied above and makes no query-error-budget claim. The horizon $H_{\max} = 4$ is frozen before calibration. A deterministic split gives 148 reliability-calibration and 148 evaluation sessions; complete-session calibration at $\delta_h = 0.05$ takes rank 142 of 148 and returns $\hat{b}_{\mathrm{cert}} = 1$, with the calibration score histogram $\{0 : 115, 1 : 30, 2 : 3\}$. Under the persistent-child panel the same rank returns $\hat{b}_{\mathrm{cert}} = 2$, which is the expected response to a heavier error regime. Structural, feedback, and joint coverage are 0.953, 0.966, and 0.932, respectively; joint coverage is the event $E_s \cap E_h$, equivalently target retention at every prefix. Table 2 reports the per-arm rows. The session identifier digest is `414a14ceae627fe3d196499bd95edd413a69d82dd9bc94c7554ca2b0a830685b`.

The binary-query integration and repeated-calibration results use separate development bundles, documented in Appendices E and F.

Structural estimand and resampling. For a fixed method, corruption rate, and metric, let $z_{ag\ell}$ be the metric for graph g from source a under corruption seed ℓ. With $D = 6$ sources, $K = 5$ seeds, and N_a graphs in source a, the frozen Table 1 estimator is

$$\widehat{\mu}_{\mathrm{macro}} = \frac{1}{D}\sum_{a=1}^{D}\frac{1}{N_a}\sum_{g=1}^{N_a}\left(\frac{1}{K}\sum_{\ell=1}^{K} z_{ag\ell}\right).$$

The test-graph counts are AGROVOC 60, ANZSRC 2, MeSH 43, OEWN noun 60, OEWN verb 14, and UNESCO 45 (224 total), as recorded in the frozen holdout manifest. Each source receives weight $1/6$, including the two-graph ANZSRC slice. This differs from weighting all 224 graphs equally. It estimates performance on these six fixed source strata, not on a random new domain. This empirical summary does not establish conditional coverage for each fit or verify the exchangeability assumptions required by the theorem. The frozen analysis averages seeds first, pairs methods by clean graph, resamples graphs within source, and recomputes the macro contrast in each bootstrap draw. Sign flips apply to the seed-averaged paired graph differences before macro aggregation. Their validity requires joint sign-exchangeability under the null, a modeling assumption rather than a consequence of randomized method assignment. Shared graphs, seeds and corruption draws establish pairing, not null sign-exchangeability. Holm adjustment covers all four rates times five planned contrasts. The source counts and weights are metadata and the released primary estimator; no new holdout evaluation or post-hoc primary-metric selection is involved.

Operational definitions of the traditional baselines. No-repair returns the observed graph, which corruption can make invalid. Local argmin chooses each non-root child's lowest raw-score parent, breaking ties by parent identifier, without a global tree constraint. Maximum arborescence minimizes the sum of these raw scores over rooted spanning trees. The minimum-edit decoder minimizes that sum plus $\lambda|T \setminus T_{\mathrm{obs}}|$, where T_{obs} is the observed edge set and λ is frozen before evaluation. Neither point decoder calibrates a set threshold. Sorted node/edge insertion and the frozen NetworkX version determine solver ties; the documented MILP fallback does not add a separate lexicographic objective. Edge-pooled conformal calibrates one quantile from pooled true-edge nonconformities. Bonferroni per-child conformal uses one score per graph and ordered child slot, at level α_s/m for the frozen family cap m. Both apply a monotone feasibility closure, adding complete tied score tiers, and certify edits relative to the observed graph.

Our finite-population learn-then-test adaptation uses each graph's any-wrong-certificate indicator. Within each domain, calibration is a simple-random sample without replacement from the fixed calibration–test population. Each finite member of the frozen threshold grid receives a one-sided exact hypergeometric bound at error level δ/M, where M counts finite thresholds. The $+\infty$ full-universe structural backbone is an analytic zero-risk fallback under fixed-root realizability and is not counted in M. Among passing thresholds, the method maximizes the mean number of certified edits on the calibration sample, with utility ties resolved toward the larger threshold. All certificate arms use the same observed graph for reference edges.

Frozen corruption and feedback generators. The sealed structural protocol applies $\max\{1, \mathrm{round}(r(|V|-1))\}$ operations at positive corruption rate r, sampling children by seed and cycling through balanced shuffled blocks of five operators: valid reparenting, orphaning, edge reversal, two-cycle injection, and a lexically hard parent swap. Valid reparenting and hard swaps exclude the child's clean subtree; every operation is retained in an audit trace. The frozen edge scorer combines word/character TF–IDF similarities and lexical features with observed-graph features, standardizes this feature vector, and fits a logistic classifier only on the frozen training graphs with six sampled negative parents per positive edge. Adaptive prediction-set nonconformity is calibrated separately by domain. The feedback study uses three simulated panel members, raw-vote error probability 0.15, two-round child blocks, and $H_{\max} = 4$. The iid regime samples member errors independently, whereas the persistent-child regime shares a child-specific misconception across repeated questions. The complete machine-readable definitions and seeds are in `configs/final_confirmatory_v5.json` and `configs/joint_e2e_public_dev_v1.json`.

B.1 Evidence scope and excluded annotation records

Previously reported expert-annotation pilot statistics and workflow-validation claims are excluded because the available provenance does not substantiate their attribution to the described human annotators. None of the retained experiment configurations binds these annotation records as training, calibration, or evaluation inputs. The records are not reclassified as synthetic data. Internal source and revision records are preserved; excluded individual responses are not included in the anonymous supplement. The reported feedback evidence remains programmed-answerer simulation, not human-interaction evidence.

B.2 Learned-scorer public-development audit

The directed GNN experiment freezes the local MiniLM snapshot, candidate construction, code inventory, public benchmark manifests, and three training seeds (13, 42, and 137). Each seed has a create-once encoder cache and an independently hashed scorer state. Aggregation first averages seeds for the same test graph, corruption rate, and method, and only then summarizes across graphs. Consequently the 296 graphs at each corruption rate, rather than seeds or edges, are the descriptive units. All 7,104 paired cells completed; the final holdout was not accessed.

The learned-scorer summaries in this paragraph weight graphs equally; Table 14 instead averages within each source and then weights sources equally. These are distinct descriptive estimands, not competing versions of one estimate. At rates 0.10/0.20/0.30, CARTER graph conformal has whole-tree coverage 0.9032/0.8885/0.8941, any-wrong-certified-edit rates 0.0158/0.0034/0.0045, and mean certified edits 0.5957/0.8412/1.2095. Pooled edge conformal has coverage 0.3198/0.2905/0.3063 and wrong-certificate rates 0.4989/0.4381/0.3446. Bonferroni child conformal has unit coverage and zero wrong certificates, with 0.4088/0.7646/1.1543 edits. Fractional graph indicators arise only from within-graph averaging of the three frozen training seeds. This development-only experiment does not establish learned-scorer performance on the sealed holdout. The aggregate result-manifest payload SHA-256 is `b99239215b225177003c3d107f86a152fce43e8701b4ede29b7f2a646a85e17f`.

B.3 Source releases

The source taxonomies come from public releases, used under their respective licenses and pinned by file name and SHA-256 in the reproducibility snapshot. The six source strata are built from Open English WordNet 2025 for the noun and verb slices (Open English WordNet Community, 2025; McCrae et al., 2020), the UNESCO Thesaurus (UNESCO, 2026), MeSH 2026 (U.S. National Library of Medicine, 2026), AGROVOC (Food and Agriculture Organization of the United Nations, 2026), and ANZSRC Fields of Research 2020 (Australian Bureau of Statistics, 2020). The temporal track additionally uses the preceding releases of MeSH and Open English WordNet, namely MeSH 2025 (U.S. National Library of Medicine, 2025) and Open English WordNet 2024 (Open English WordNet Community, 2024), so that a transition is a genuine editorial change rather than a simulated one. The learned scorer of Appendix B.2 and the temporal track embed node text using the pinned local sentence-transformers checkpoint `all-MiniLM-L6-v2`, which is based on MiniLM (Wang et al., 2020). The AGROVOC extraction uses English labels; FAO's official FAQ licenses English and its other five official-language versions under CC BY IGO 3.0, while other languages retain their providers' terms. The release manifest records the project's downloaded snapshots; a subsequently updated resource web page does not redefine those files.

B.4 Source-stratified calibration diagnostic

Split construction. The reliability-calibration and evaluation halves are separated without consulting any label. Within each source, instances are ordered by a content hash and halved at the midpoint, which stratifies both halves across all six sources (30/30 for AGROVOC, MeSH, OEWN noun and OEWN verb, 17/17 for UNESCO, 11/11 for ANZSRC) and makes the assignment reproducible from the instance identifiers alone. The ambiguity-block query policy, its lexical tie-break, block size two, and $H_{\max} = 4$ are fixed in the run configuration before any session is generated, and every paired arm replays one frozen transcript, so the arms differ only in how they treat the answers they receive.

Theorem 1 assumes that complete sessions are exchangeable under one frozen pipeline. Exchangeability cannot be established from a finite diagnostic. Table 5 instead checks visible source heterogeneity in the canonical feedback run by splitting the 148 reliability-calibration sessions by source. The calibration score is concentrated near zero in every source, with per-source means between 0.13 and 0.33 and no observed count above two. Recalibrating with each source removed returns the same budget in all six cases, so the selected order statistic is not driven by any one observed source. These checks support, but do not verify, the exchangeability assumption.

Evaluation-side joint coverage varies more widely, from 0.818 to 1.000. The low end comes from ANZSRC with 11 sessions and AGROVOC with 30. The two smallest slices are ANZSRC and UNESCO (17 sessions); their different observed coverage illustrates the limits of interpreting small source slices. This variation is consistent with evaluating a marginal guarantee on small slices, but it is not a conditional coverage assessment, which the theorem does not provide. Lifetime error remains between 0.000 and 0.033 everywhere, and the hard filter errs more than the calibrated budget in every source.

Table 5: Source-stratified calibration diagnostic on the canonical feedback run (iid panel, ambiguity policy). "Mean excl." and "max" summarize target-parent exclusion counts on the calibration half; "b w/o source" recalibrates at $\delta_h = 0.05$ with that source removed; coverage and error are on the evaluation half.

	Calibration half				Evaluation half		
Source	n	Mean excl.	Max	b w/o source	Joint cov.	Wrong	Hard b=0
AGROVOC	30	0.333	2	1	0.833	0.000	0.133
ANZSRC FoR	11	0.273	2	1	0.818	0.000	0.091
MeSH	30	0.233	2	1	0.967	0.033	0.067
OEWN noun	30	0.300	1	1	1.000	0.000	0.033
OEWN verb	30	0.133	1	1	0.933	0.033	0.067
UNESCO	17	0.176	1	1	1.000	0.000	0.118
All sources	148	0.243	2	1	0.932	0.014	0.081

The $2/148$ lifetime wrong-edit count for the all-source calibrated arm can also be summarized by a one-sided 95% Clopper–Pearson reference bound of 0.042 under iid Bernoulli sessions (Clopper & Pearson, 1934). That reference bound is distinct from the paper's unconditional marginal population guarantee, and from the fixed-source programmed-panel diagnostic reported in Table 5.

We do not audit per-answerer variation, because the answers in this study come from a programmed panel rather than identified individuals. Measuring human answerer heterogeneity would require separately documented participants, session-level calibration, and an appropriate human-study protocol.

B.5 SAFETY–UTILITY OPERATING POINTS AND RISK SWEEP

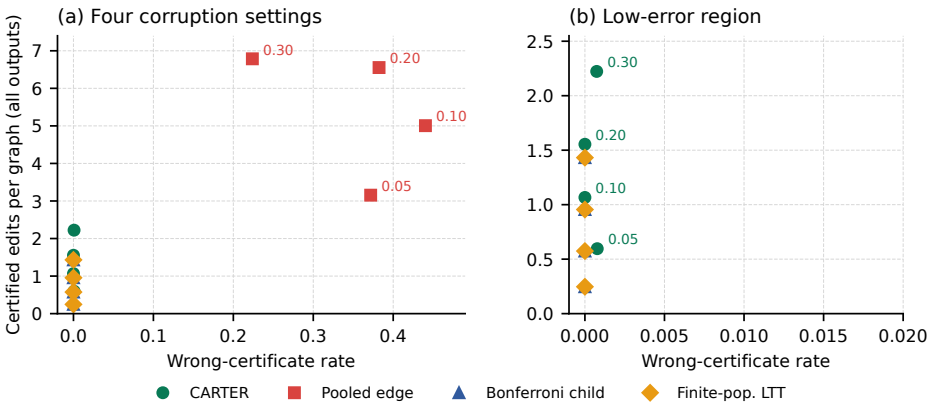


Figure 2: Sealed-holdout safety–utility operating points. Labels are corruption rates at the fixed structural level $\alpha_s = 0.10$. The right panel magnifies the safe region; Bonferroni per-child and finite-population LTT coincide. Certified-edit counts are output quantities to interpret together with the wrong-certificate rate.

The sealed table fixes one risk level, $\alpha_s = 0.10$, so a reader cannot tell whether the ordering it reports is a property of the methods or of that particular level. Because the one-shot evaluation has already been used, we answer this on development data instead, sweeping the nominal structural level over $\alpha_s \in \{0.02, 0.05, 0.10, 0.20, 0.30\}$ and rerunning the whole procedure at each value. Every method sees the same graphs, corruption draws, seed, and solver; only α_s changes. Figure 3 plots the resulting risk sweep and Table 6 lists the coordinates. This TF-IDF sweep fixes seed 13 and corruption rate 0.20. It averages within each of four sources (MeSH, OEWN noun, OEWN verb, and UNESCO; 60/60/60/34 graphs), then weights the four source means equally. This descriptive sensitivity analysis attaches no uncertainty interval or significance claim to the sweep and is distinct from the learned-scorer audit above.

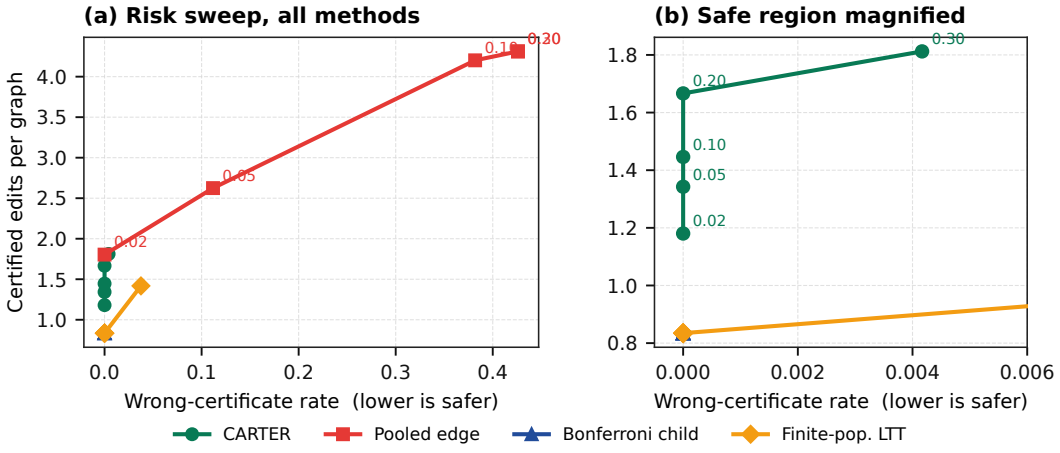


Figure 3: Development risk sweep: four-source macro averages (MeSH, OEWN noun, OEWN verb, UNESCO), TF-IDF seed 13, corruption rate 0.20. Labels are nominal levels α_s. The right panel magnifies the safe region, where Bonferroni per-child is a single point because its certificate does not depend on α_s.

Table 6: Development risk sweep, using the same four-source macro average, seed 13, and corruption rate 0.20 as Figure 3. Coverage is whole-tree, "wrong" is the graph-level wrong-certificate rate, and "edits" is certified edits per graph. Bonferroni is constant on this grid; LTT coincides with it below $\alpha_s = 0.30$ but differs at 0.30. The two methods are reported separately.

	CARTER			Edge-pooled		
α_s	Cov.	Wrong	Edits	Cov.	Wrong	Edits
0.02	0.992	0.000	1.180	0.687	0.000	1.804
0.05	0.952	0.000	1.343	0.435	0.112	2.625
0.10	0.897	0.000	1.446	0.242	0.382	4.201
0.20	0.773	0.000	1.666	0.242	0.426	4.312
0.30	0.682	0.004	1.812	0.242	0.426	4.312
	Bonferroni per-child			Finite-population LTT		
α_s	Cov.	Wrong	Edits	Cov.	Wrong	Edits
0.02	1.000	0.000	0.835	1.000	0.000	0.835
0.05	1.000	0.000	0.835	1.000	0.000	0.835
0.10	1.000	0.000	0.835	1.000	0.000	0.835
0.20	1.000	0.000	0.835	1.000	0.000	0.835
0.30	1.000	0.000	0.835	0.862	0.038	1.418

At $\alpha_s = 0.10, 0.20, 0.30$, realized four-source macro coverage falls below $1 - \alpha_s$. Marginal validity does not require every fitted split or fixed-source summary to exceed its nominal level. This single-seed sweep does not distinguish sampling variability from departures from the structural-calibration assumptions. Appendix F instead repeats feedback-budget calibration, not this structural experiment.

Three readings follow. First, the separation is not an artefact of one level: at every α_s from 0.05 upward CARTER certifies with a wrong rate of at most 0.004 while edge pooling ranges from 0.112 to 0.426, and the gap widens before the discrete outputs plateau. Second, pooled coverage is 0.242 from $\alpha_s = 0.10$, whereas its wrong rate rises from 0.382 at 0.10 to 0.426 at 0.20 and stays there at 0.30. Larger risk levels lower the nonconformity cutoff; finite candidate sets can leave the resulting certificates unchanged between levels. Third, at $\alpha_s = 0.02$ both methods have zero observed wrong certificates and edge pooling certifies more (1.804 against 1.180), but its whole-tree coverage is lower (0.687 against 0.992). Equality of the observed output-error rate therefore does not make the structural coverage guarantees or deployment requirements equivalent.

B.6 CONFORMAL RISK CONTROL ON THE SAME BUDGET

Conformal risk control (CRC) (Angelopoulos et al., 2024) assumes exchangeable, bounded loss functions that are nonincreasing in their parameter *for each sample*, together with its endpoint and right-continuity conditions. The finite-sample selection criterion is $\frac{n}{n+1}\hat{R}_n(b) + \frac{1}{n+1} \leq \delta_h$ for a binary loss; it is not a confidence bound on the observed empirical risk. Our original adaptation applies this criterion to the indicator of any wrong certified edit over a complete session, taking the smallest qualifying integer budget. Table 7 preserves its measured results.

Correction to the applicability claim. The original check verified only that the *average* calibration loss decreased with budget. This does not establish CRC's pointwise condition. In the saved development evaluation, the session `mesh-00022` under the independent-panel, ambiguity-block policy has lifetime losses $0, 1, 0$ at budgets $0, 1, 2$. At a small budget an empty version space suppresses output; a larger budget can restore a nonempty space with a wrong certificate. Thus the raw adapted loss is not nonincreasing on its domain. We withdraw the claim that the published monotone-loss guarantee applies to this arm, and label it a raw CRC-rule heuristic. This correction concerns our adaptation, not the validity of CRC's theorem. The original per-session calibration losses were not saved, so an aggregate-only artifact cannot establish their pointwise behavior.

Table 7: Calibration risk $\hat{R}_n(b)$ on the reliability half, and the budget each principle selects. The raw CRC-rule heuristic and CARTER select the same budget in both regimes. The tie is retained; the heuristic does not inherit CRC's monotone-loss guarantee.

	$\hat{R}_n(b)$					selected b	
Regime	0	1	2	3	4	Raw CRC	CARTER
Independent panel	.101	.000	.000	.000	.000	1	1
Persistent child error	.223	.223	.000	.000	.000	2	2

The unrounded budgets are one and two, respectively, and the selected arms match on every saved evaluation-session metric in each regime, excluding the accounting-only allocation of shared runtime. Both methods use the same candidate graph, transcript, certificate solver, and budget; hence their version spaces and certificate sets are identical by construction, not merely after rounding. This identifies the shared-budget state identity; empirical comparisons depend on the surrounding interface and calibration choices.

A conservative repair and its scope. For a fixed finite budget grid define $\overline{L}_i(b) = \max_{k \geq b} L_i(k)$, taking the maximum *within each session before averaging*. This envelope upper-bounds the original loss and is nonincreasing. An additional output-abstaining endpoint has loss zero, so the endpoint condition can be satisfied even when structural errors survive the largest numerical budget. CRC can be applied to this envelope under its other assumptions. The monotone-envelope construction is CRC's existing extension to nonmonotone losses (Angelopoulos et al., 2024, Section 2.3, Corollary 1), not a new calibration theorem here. We implement this as a separate development-only diagnostic, not as a retrospective change to the historical raw row of Table 2. CRC can also calibrate a state-containment loss; we do not claim that it is inherently unable to protect state.

A derived envelope comparator on the original development data. The calibration means identify the envelope's selected budget without reconstructing unobserved per-session losses. The independent-panel wrong counts over budgets $0, \ldots, 4$ are $(15, 0, 0, 0, 0)$ out of 148; under persistent errors they are $(33, 33, 0, 0, 0)$. Because the losses are nonnegative, every per-session loss in the zero-count suffix is zero. The envelope therefore has corrected empirical criterion $1/149$ from budget one or two, respectively. At every smaller budget its criterion is at least the raw value, $16/149$ or $34/149$, exceeding 0.05. Thus the pointwise-envelope rule, with the formally included all-abstain endpoint, selects exactly the same budgets as CARTER. The persistent-regime envelope risk at budget zero is *not* identified from the two marginal counts; only the interval $[33/148, 66/148]$ follows. The supplement verifies these bounds with exact rational arithmetic and checks the two saved calibration artifacts against each other. The added envelope row in Table 2 is a post-hoc development-derived comparator: it reuses the same-budget outputs by deterministic identity. It is neither an independently executed evaluation nor a confirmatory result. Its formal guarantee remains conditional on CRC's frozen-function and exchangeability assumptions. The corrected comparator has the same selected budgets and outputs as CARTER.

State versus output: an existence result. An exhaustive four-node example uses root 0, fixed edges $(0, 2), (0, 3)$, and three possible parents $0, 2, 3$ for node 1. The target uses parent 0, whereas the observed tree uses parent 2. A budget-zero answer allowing only $\{2, 3\}$ removes the target but leaves two trees, no conflict, and no certified edit relative to the observed tree. Enumeration verifies that zero wrong-output loss need not imply target containment. It does not demonstrate a real-data performance difference between equal-budget arms.

Three safety events. Let S be the event that the canonical target is absent from the maintained family at some prefix, G the event that some reported mandatory or impossible edge statement (including exclusions outside C) is false at a prefix, and O the event that an edit derived from these statements is wrong. At conflict, all operational certificates are withheld. For the exact certificate interface,

$$\mathsf{O} \subseteq \mathsf{G} \subseteq \mathsf{S}.$$

Indeed, every executed addition or deletion is a reported edge statement. If the target belongs to a nonempty family, all intersections and complements of unions are correct relative to that target. These implications hold prefix by prefix and hence after taking the union over the session. Their converses need not hold. In the four-node example above, the target edge $(0, 1)$ becomes impossible although it is absent from the observed tree; the surviving alternatives yield no addition. Thus S and G occur, but O does not. This establishes a distinction between interfaces, not an empirical advantage over an output-risk controller. Output-risk control can suffice when an application uses only the certified edits.

A binary CRC special case. Suppose the complete-session scores $Y_1, \dots, Y_n$ are fixed integers in $\{0, \dots, H_{\max}\}$. The budget must not change the query policy, stopping rule, answer aggregation, or randomization that generates these scores. Use the full integer budget grid and the strict-exceedance loss $\ell_i(b) = \mathbf{1}\{Y_i > b\}$. For the same calibration sample and $0 < \delta_h < 1$, the binary CRC selection criterion (Angelopoulos et al., 2024) satisfies the exact identity

$$\frac{\sum_{i=1}^{n} \mathbf{1}\{Y_i > b\} + 1}{n+1} \le \delta_h \quad \Longleftrightarrow \quad \#\{i : Y_i \le b\} \ge k := \lceil (n+1)(1-\delta_h) \rceil .$$

For $k \le n$, the smallest feasible budget is therefore $Y_{(k)}$, including ties. For $k = n + 1$, the empirical feasible set is empty. The prescribed fallback $H_{\max}$ has zero feedback-exceedance loss because $Y \le H_{\max}$ for every session; it does not satisfy the empty empirical criterion. This is a special-case identity, not a new calibration principle. It concerns feedback-budget exceedance, not state loss caused by initial structural exclusion, and does not equate general output-CRC with our method. The associated marginal probability guarantee additionally requires the stated exchangeability of calibration and deployment sessions under the same frozen pipeline.

Maximum budget and the output-abstaining endpoint. Since each answer contributes at most one unit, $L_t(T) \le t \le H_{\max}$ for every $T \in \mathcal{V}_0$. Hence $\mathcal{V}_t^{H_{\max}} = \mathcal{V}_0$: feedback adds no further restriction, but initial structural certificates remain available. This fallback differs from the output-abstaining endpoint of the output-loss envelope, which withholds all edits and has zero output loss by definition, without implying target containment. On a frozen finite budget grid, the envelope $\overline{L}_i(b) = \max_{j \ge b} L_i(j)$ must be formed for each session before averaging. It is a nonincreasing upper bound; adding a zero-loss abstaining endpoint preserves all finite-budget envelope values. Its risk guarantee retains the other conditions of the invoked CRC theorem.

Scope of the CRC applicability correction. Our raw output-loss adaptation does not satisfy the pointwise monotonicity condition of the cited 2024 CRC theorem. This statement is limited to that adaptation and theorem: later work studies risk control for nonmonotonic losses through algorithmic stability (Angelopoulos, 2026). We do not establish those stability conditions for the present adaptation.

Development-only diagnostic grid. We additionally enumerate all rooted trees on four nodes in a complete 2^6 synthetic factorial, with 80 calibration and 120 evaluation sessions per cell. The factors are a per-child candidate-omission generator parameter $(0.05, 0.15)$, feedback calibration level $(0.05, 0.15)$, candidate-recall generator parameter $(0.8, 1)$, persistent-error mixture $(0, 0.8)$, session length $(2, 4)$, and sampled candidate ambiguity $(2, 3)$. The omission knob is *not* a calibrated structural miscoverage level; realized recall and fallback to a feasible candidate graph are logged. All arms share the generated candidates, observed tree, answer transcript, and seed; this toy mechanism experiment uses no learned scorer or human answerer. Raw lifetime loss is pointwise nonmonotone on calibration samples in 53 of the 64 cells. The other cells do not prove global monotonicity. The raw rule, pointwise-envelope rule, and disagreement order statistic are reported in full, including target containment at every prefix, first target loss and wrong output, silent state failure, conflict, and certified edits. Paired bootstrap intervals are descriptive and conditional on the selected calibration budgets; simultaneous Hoeffding bounds cover the prespecified family of 512 normalized competitor contrasts. No significance, equivalence, or real-data superiority claim is made. The supplement contains all 23,040 method-session rows and 92,160 selected-arm prefix records, configurations, code, and checksums. Enumerated witness availability is an existence check, distinct from materialized solver-witness validation in the scaling audit.

B.7 NOISE-ROBUST ONLINE CONFORMAL PREDICTION AS A COMPARISON ARM

The historical adapter keeps the hard filter but adjusts its threshold using the noise-robust online conformal update of Xi et al. (2025). Its saved outputs use the same frozen graphs, transcripts, and certificate code as Table 2; the reconstruction limits are stated below.

Adaptation and what it preserves. Their method calibrates a threshold for online *label* prediction under uniform label noise of known rate, and its coverage guarantee is for that setting; it does not transfer to lifetime version-space containment, and we claim no such property for it here. We map a query to the input, the candidate parents of the queried child to the classes, the edge score to the nonconformity score, and the possibly wrong answer to the noisy label. The threshold starts at the structural threshold and is updated after each answer by their robust pinball gradient. Two variants are reported because the choice is ours, not theirs: *set only* passes the conformal parent set as the constraint, and + *answer* lets the answer narrow that set when it falls inside it. The baseline is given the true noise rate ($\epsilon = 0.15$) that its update rule requires and that a deployment would have to estimate; CARTER needs no per-answer noise model at all, since it calibrates a quantile of complete-session exclusion counts. The baseline's learning rate is chosen on the reliability half by minimizing the absolute difference between whole-tree every-prefix retention and $1 - \alpha$, with larger mean certified edits breaking ties. Over $\eta \in \{.005, .01, .02, .05, .1, .2\}$, the selected values are .01 and .05, respectively. This is our session-level adapter selection rule, not the original method's per-round label-coverage objective. Replaying the *no feedback* and hard-filter arms through this code reproduces their Table 2 entries exactly. This validates the shared certificate replay path, not the NR-OCP adapter's score reconstruction.

Saved results and provenance. The set-only adapter matches the no-feedback certified-edit mean of 1.345; the answer-narrowing adapter has 0.068 lifetime error, 0.791 retention, and 3.088 certified edits per session. These historical, development-only session summaries are retained under the adapter's actual selection rule. The historical artifacts do not contain the per-round prediction sets or sufficient statistics for the previously reported set widths and label coverage. We withdraw those diagnostics and their mechanism interpretation. A fixed-parameter score reconstruction failed byte-identity checks; the magnitude and decision-level effect of the score differences remain unresolved. Thus these rows describe our adapters, not a verified superiority result over the original NR-OCP method.

B.8 ANNOTATOR-RELIABILITY AGGREGATION

Reliability-weighted truth inference is the classical alternative to plurality: estimate each answerer's accuracy and aggregate votes by posterior weight (Dawid & Skene, 1979). We implement the one-coin Dawid–Skene model on the frozen transcripts, whose three raw panel votes per event are recorded per member. EM estimates the per-member reliabilities *without ground truth* on the reliability half; on the evaluation half, we then replace each event's answer with the posterior-argmax vote and apply it as a hard constraint, making this the reliability-weighted counterpart of running plurality, replayed through the same certificate code.

Under independent errors, EM recovers the true reliabilities well ($\hat{p} = 0.84$–0.90 against a true 0.85), but with three near-equal weights the posterior argmax reduces to the per-round majority, and the arm lands exactly on the hard filter: 0.081 lifetime error, 0.736 retention. This arm is therefore not the running-plurality arm of Table 2, and the two differ on both axes: Dawid–Skene aggregates the three raw member votes *within* a round and emits one constraint per round, whereas running plurality accumulates the aggregated panel answers *across* rounds and emits a single constraint per child. That is why their numbers differ even though both are majority rules. Under persistent errors the failure is sharper and structural: because all three members share the same misconception, votes are unanimous, and EM estimates $\hat{p} = 1.000$ for every member—the model is maximally confident precisely when the panel is most wrong, since perfectly correlated error is unidentifiable from agreement. The arm then errs on 0.216 of sessions, again matching the hard filter, whereas the calibrated budget errs on 0.000. Reliability weighting addresses *who* is wrong; the budget addresses *how often anyone* is wrong, which is the quantity that remains estimable under shared error.

B.9 CANDIDATE-RECALL STRESS TEST

The main experiments verify $T^\star \in \mathcal{T}((V, C), r)$ by construction, which isolates the calibration layer. This study measures the failure *mode* when that precondition breaks: for a controlled fraction of non-root children, the true parent edge is removed from the candidate set, and the frozen sessions are replayed at the calibrated budget (Table 8).

Table 8: Candidate-recall stress test (148 evaluation sessions, independent panel, $\hat{b}_{\mathrm{cert}} = 1$). The first column is the nominal recall setting, not realized recall. Coverage is initial structural containment; it is not every-prefix target retention. Conflict is the historical reported flag, which does not separately log solver exceptions.

Setting	Coverage	Wrong	Conflict	Edits
1.00	0.953	0.014	0.027	3.439
0.95	0.041	0.419	0.574	1.689
0.90	0.000	0.399	0.696	1.358
0.80	0.000	0.351	0.791	1.264

Degradation is not confined to abstention. Once the true parent is missing, the version space consists of trees that are all wrong about that child, and they can agree: at nominal recall 0.95 the wrong-certificate session rate rises to 0.419 even though conflict also rises. This illustrates the necessity of candidate inclusion—the observed failure lies outside the theorem—and makes the precondition operational: a deployment must audit generator recall separately, because no calibration downstream of a recall failure restores safety. This differs from the weak-*scorer* case of Appendix B.11, where candidates were complete and the method degraded to abstention. The generator rounds the number of removals within each graph. At setting 0.95 it removes 142 of 2,431 true-parent candidates: realized child-weighted recall is 0.941588 and graph-mean recall is 0.939563. The historical replay catches solver exceptions as conflict, so its saved conflict counts cannot separate infeasibility from execution errors; no unrecorded exception breakdown is inferred here.

B.10 ANSWER REGIMES BEYOND INDEPENDENT NOISE

The paired study fixes one answer regime. A separate controlled study compares five answer regimes, on 36 synthetic rooted graphs per cell with the same six arms and a frozen five-round transcript, so that the arms differ only in how they treat the answers. Beyond independent noise and the persistent child errors of the main text, it adds *ambiguity-targeted* errors, which concentrate on the concepts a policy is most likely to ask about; *set-valued* replies, which return several acceptable parents at once; and *abstaining* replies, which consume a question without constraining anything.

Table 9 reports terminal-round edit counts together with lifetime safety columns. In these cells the hard filter loses the target in 50–81% of runs and certifies a wrong edit in 42–69% of runs, in every regime, while a budget of one holds wrong edits at ≤ 0.111 and a budget of two has no observed wrong edits. Ambiguity-targeted and persistent errors produce the largest hard-filter error rates in this grid. Set-valued replies are the one regime where the hard filter certifies markedly less (1.31 edits versus 3.44–4.08 in the other regimes), because a wider acceptable set removes fewer structures per answer.

Table 9: Five answer regimes, terminal-round edit counts, ambiguity-directed policy, 36 graphs per cell. "Wrong" is the lifetime wrong-certified-edit rate and "retained" is the fraction of runs keeping the target through every round. Fixed budgets are used here because this controlled study has no calibration stream of its own.

	Hard b=0		Budget b=1		Budget b=2	
Answer regime	Wrong	Retained	Wrong	Retained	Wrong	Retained
Independent	0.583	0.389	0.000	0.750	0.000	1.000
Persistent child	0.694	0.194	0.111	0.667	0.000	0.917
Ambiguity-targeted	0.667	0.222	0.111	0.583	0.000	0.861
Set-valued	0.472	0.361	0.000	0.667	0.000	0.917
Abstaining	0.417	0.500	0.028	0.833	0.000	0.972

B.11 GENUINE TEMPORAL TRANSITIONS

Three leakage-audited MiniLM runs are averaged within each of six old-to-new MeSH/OEWN test instances; old-release text and old primary-parent edges are the only model-visible inputs, and new-release text is never exposed. Table 10 reports the seed-averaged aggregate. The unchanged old tree is the strongest point predictor. For certificate methods, we separately decode an auxiliary score-plus-edit-penalty arborescence within the method's allowed graph, using the frozen edit penalty of 1. CARTER's auxiliary point metrics coincide with minimum-edit on all six instances, while its unanimous certificate emits no edits. Its zero wrong-certificate rate therefore reflects abstention, not accurate committed repairs. Edge-pooled calibration has mean error frequency 0.667 over the same six instances (four errors per seed on average); all six incur an error in at least one seed. With two release-pair clusters and five calibration graphs this track is descriptive and supports no coverage claim.

Table 10: Genuine MeSH/OEWN release transitions, three MiniLM seeds averaged within each of six test instances. Parent accuracy and "Changed-only" evaluate point predictions (auxiliary decoders for certificate methods); Coverage, Wrong, and Edits evaluate set certificates. "Changed-only" is restricted to nodes whose primary parent actually changed between releases.

Method	Parent acc.	Changed-only	Coverage	Wrong	Edits
No repair	0.923	0.000	—	—	—
Local argmin	0.352	0.444	—	—	—
Maximum arborescence	0.371	0.333	—	—	—
Minimum-edit arborescence	0.683	0.167	—	—	—
CARTER	0.683	0.167	1.000	0.000	0.000
Edge-pooled conformal	0.671	0.167	0.333	0.667	4.556

B.12 COMPILER AND CERTIFICATE SCALING

The frozen reverse-adjoint study used generated graphs only, CPU-isolated workers, and one thread per numerical library. Its grid has two seeds, sparse and dense regimes, and four query budgets $b_{\mathrm{qry}} = 0, 1, 2, 3$. All 160/160 planned algorithm cells completed without a guard skip, work-cap skip, timeout, or worker error. On the overlap sizes $n \in \{8, 12, 17\}$, all 48/48 comparable reference–adjoint pairs have identical hashes for both the complete state/marginal coefficient payload and the selected-query records at horizons 1, 3, 6, and 12. Thus the parity claim is exact; it is not a floating-point tolerance check.

Each cell has one `perf_counter` timing observation. The descriptive median ratio of per-edge reference time to reverse-adjoint time is $8.002\times$ over all 48 pairs (Figure 4). The adjoint-only grid also completes at $n = 48$, with $|D_0| = 846$ in the dense regime and 188 in the sparse regime; the largest observed compile time is 9.55 seconds (the v2b adjoint artifact underlying Figure 4). These single runs do not quantify repeated-run uncertainty. The separate E1 study in Appendix E uses 36 inputs and three timing repetitions per input; its $2.407\times$ compute-plus-serialization ratio describes a different grid and timing unit.

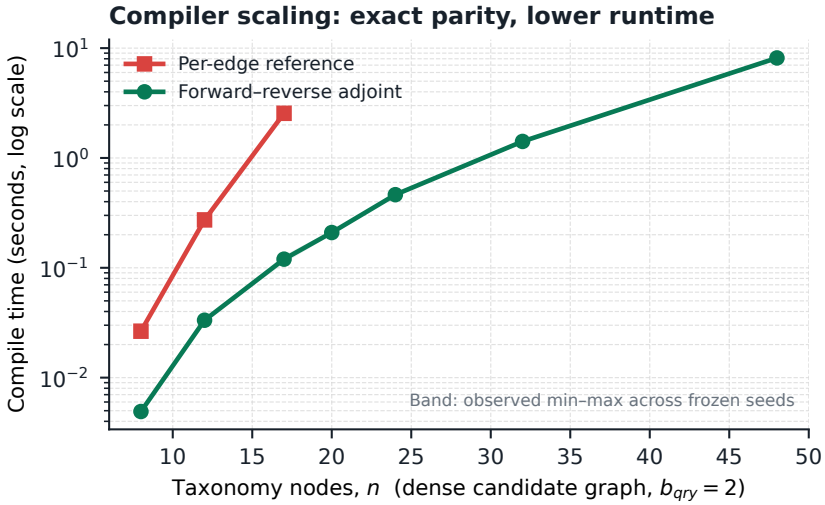


Figure 4: Dense-regime compiler scaling at $b_{\mathrm{qry}} = 2$. Points are medians over the frozen seeds and bands span their observed minimum–maximum. The reference is intentionally guarded beyond $n = 17$; the exactly matching forward–reverse implementation completes through $n = 48$.

The implementation stores a straightforward forward and reverse tape. At $n = 48, b_{\mathrm{qry}} = 3$, this proxy reaches about 19.0 million coefficient slots, while sampled whole-worker RSS spans approximately 0.91–1.03 GiB across the sparse and dense regimes. RSS was polled every 10 ms by the parent process and is platform- and allocator-specific; it is neither an exact peak nor a deterministic memory bound. Likewise, the complexity statement counts arithmetic operations in the truncated integer-polynomial ring and excludes exact-integer bit growth. These synthetic, secondary development results do not establish production scalability. The frozen configuration payload SHA-256 is `833c76d5265ed4df6cf716da265383c1296b7285da54844a1b6fd76fca35a99d`, and the validated artifact-manifest payload SHA-256 is `6bcc64a9c0ea78dd8883086debc887a2306dbfc7e720dd01ee957518d470775c`.

An earlier v1c bundle remains as a transparent per-edge reference-compiler and force/forbid-certificate boundary. It planned 128 cases and retained all outcomes: 106 completed, 16 were rejected by the predeclared $n > 18$ reference-compiler guard, two dense $n = 17, b_{\mathrm{qry}} = 3$ cases exceeded the 10^8-operation preflight cap, and four dense $n = 48, |D_0| = 846$ certificate cases exceeded the 650-edge cap. There were no timeouts or worker errors. The same CSV column stores b_{qry} for compiler cases and b_{cert} for certificate cases; these are separate factors.

Table 11: Representative completed v1c reference and certificate cells. Ranges span the frozen seeds or budgets stated in the row; "calls" are determinant calls for compilation and arborescence-oracle calls for certificates.

Operation/regime	n	$\lvert D_0\rvert$	Budget	Calls	Time
Per-edge reference compile, dense	17	112	$b_{\mathrm{qry}} = 2$	113	2.52–2.66 s
Certificate, dense	36	455	$b_{\mathrm{cert}} = 3$	456	2.39 s
Certificate, sparse	48	188	$b_{\mathrm{cert}} \in \{0, 1, 2, 3\}$	189	0.645–0.680 s

The maximum sampled whole-worker RSS in v1c was 50.75 MiB, but allocator and platform effects make it non-reproducible as a bound. The artifact therefore also records deterministic retained-integer, serialized-JSON, coefficient-slot, and reference-byte proxies. The retained per-edge reference compiler—not the current query compiler—performs one division-free determinant for Z and one for every allowed edge, giving $O((|D_0| + 1)n^4(b_{\mathrm{qry}} + 1)^2)$ arithmetic operations for that reference implementation. The current compiler instead uses the single forward–reverse determinant circuit of Theorem 3. The certificate code performs at most $|D_0| + 1$ force/forbid

oracle calls and uses NetworkX as a transparent reference solver. None of these measurements is presented as a production-scale result. The validated v1c artifact-manifest payload SHA-256 is `1f6746ac9b999efe7fb5502a4faf41a6df632fd2528a38f476b26c7f8a2a7fa9`.

Cost of the budgeted certificate. We next measure feedback-dependent certification separately from query compilation. The sealed holdout certifies without feedback and therefore exercises the structural path, whose per-instance cost is a median of 2.6 ms and a maximum of 84 ms over 4,480 measurements. Once answers arrive, Theorem 2 issues at most $|D_0| + 1$ force/forbid solves, which is a different cost. Table 12 times that path directly on generated rooted taxonomies with k candidate parents per non-root node, one wrong answer, and $b_{\mathrm{cert}} = 1$, using the same NetworkX reference solver. Cost grows roughly linearly in $|D_0|$ and superlinearly in n: about one second at $n = 50$ with 240 candidate edges, ten seconds once the candidate set reaches a thousand edges, and 152 seconds at $n = 500$ with the sparsest candidate set evaluated. Thus a release-scale taxonomy lies well outside interactive latency for this reference implementation. These legacy single timings precede the separate exact-solver upgrade below. They describe this implementation and should not be extrapolated to other solvers or to release-scale taxonomies.

Table 12: Budgeted-certificate wall-clock time (development measurement, NetworkX reference solver, single timings, witnesses not materialized). Ranges span three seeds. Cells whose smaller k already exceeded 90 seconds were not run and are marked —.

Nodes n	$k = 5$	$k = 10$	$k = 20$
50	0.93–1.18 s	2.79–3.72 s	9.55–9.68 s
100	3.86–4.02 s	15.3–30.2 s	42.0–61.6 s
200	23.8–78.0 s	112 s	—
350	97.7 s	—	—
500	152 s	—	—

One caveat limits the interpretation of Table 12. Its grid places $k \in \{5, 10, 20\}$ candidate parents on each node, whereas the reported experiments use the complete within-graph parent rule, which on a 50-node unit admits up to 49 candidates per node. The tabulated times consequently do not establish latency at the density used in the paper's main evaluation. The 152-second entry at 500 nodes is a sparse-case measurement, not a worst-case runtime bound.

Exact engineering upgrade (development-only). We retain the reference implementation and provide a separate optimized module. It caches candidate order, integer-scaled weights, and incoming edge lists within each certificate call. Under each force/forbid restriction, it first chooses the unique minimum-weight incoming edge for each non-root node. If these choices form a rooted spanning tree, they achieve the sum of the independent per-child lower bounds and are the unique global optimum; hence they reproduce the reference tree, not only its objective. If they contain a cycle, the code falls back to the same NetworkX branching routine with the same node/edge insertion order and weights. Missing incoming edges establish infeasibility. Independent restricted solves can also run in a process pool; output assembly preserves reference order. This is a semantics-preserving implementation change, not a new minimum-arborescence theorem.

Table 13: New development-only certificate matrix with witnesses enabled. Cells report median method runtime in seconds and completed/attempted runs across three fixed seeds. A 30-second whole-process limit includes setup and validation; — means no completed run, not zero time. IQRs, sampled process-tree RSS, logical/actual solve counts, case hashes, and every timeout are in the supplement. The two optimized columns use one and four processes, not Python threads.

n	k	Reference	Optimized, 1	Optimized, 4
50	5	1.33 (3/3)	1.20 (3/3)	4.85 (3/3)
50	10	3.34 (3/3)	0.21 (3/3)	4.23 (2/3)
100	5	5.63 (3/3)	0.33 (3/3)	5.51 (3/3)
100	10	21.62 (1/3)	18.02 (1/3)	15.57 (3/3)
200	5	— (0/3)	— (0/3)	19.78 (2/3)
200	10	— (0/3)	2.08 (2/3)	6.22 (2/3)
350	5	— (0/3)	— (0/3)	— (0/3)
350	10	— (0/3)	4.98 (1/3)	15.77 (1/3)
500	5	— (0/3)	3.58 (3/3)	18.64 (3/3)
500	10	— (0/3)	19.80 (1/3)	— (0/3)
1000	5	— (0/3)	18.60 (1/3)	— (0/3)
1000	10	— (0/3)	— (0/3)	— (0/3)

These runs use an Intel Core i7-14650HX host, Python 3.11.9 and NetworkX 3.6.1, with fresh subprocesses and deterministic shuffled variant order. The machine is not an isolated benchmarking host. RSS is sampled every 50 ms across the worker process tree and may double-count shared pages. Completed-only medians are not estimates of uncensored median latency; in particular, a faster median based on one completed seed cannot establish an advantage over a method completing a different set of seeds. The cyclic fallbacks still dominate difficult cases, and four processes are often slower than one because of startup, copying and interprocess transfer.

The parity suite compares complete outputs, including objective values, all edge sets, baseline and witness edge identities, restricted optima, infeasibility and deterministic ties. Its exhaustive component spans 531 candidate subgraphs through four nodes and 4,199 budget-expanded reference/optimized comparisons. Random and targeted adversarial cases supplement this finite check. Timed instances are compared only when the reference and optimized outputs both complete; censored comparisons are explicitly unresolved. Witness-free runs measure the separate API mode. The profiled witness-materialization timer measures wrapper construction only: alternate trees are computed by both modes, so it is not the full cost of witness inference. Detailed parity, certificate-only, witness-mode, and longer-reference results are provided in the supplement.

The contribution is the exact structured certificate interface and its optional one-step compiler within the rooted single-parent setting. The Matrix–Tree, Berkowitz, Berlekamp-weight, and minimum-arborescence components are reused and specialized here; globally optimal adaptive querying, exhaustive polyhierarchy certification, and unconditional logarithmic query bounds remain outside this paper's scope.

C SEALED EVALUATION: FINAL RESULTS AND EVALUATION HISTORY

The identifiers v2–v5 refer to frozen benchmark and run specifications, not CARTER variants or manuscript revisions. Of these four constructions, only the final evaluation (v5) has an internally validated confirmatory receipt. The earlier invalidated and superseded constructions are disclosed below; they supply no additional confirmatory evidence.

Final structural evaluation (run ID v5). This is the internally frozen evaluation reported in the main text. Its construction excludes every node appearing in any development benchmark or in the invalidated v2 holdout, which exhausts the large-unit pools of three domains; its frozen specification therefore reduces unit sizes to 20/12/10 target nodes for OEWN verb, UNESCO, and ANZSRC while keeping 50-node units for AGROVOC, MeSH, and OEWN noun. Raw eligible pools contain 120/4/85/120/27/90 units, and the prespecified global permutation yields 222 calibration and 224 test graphs. The gate history is frozen–unlocked–claimed–verified–consumed with a validated verification record; the run specification hash, benchmark manifest hash, verification record, and frozen analysis script are included in the release package. Table 14 gives a descriptive cross-setting comparison against this holdout.

Table 14: Descriptive cross-setting comparison for CARTER. Development uses the three-seed GNN collection; the sealed test column uses the frozen TF–IDF scorer on the final six-domain holdout. Differences therefore combine changes in scorer and graph collection and are not exchangeable paired effects.

Corruption	Metric	Development (GNN)	Sealed test (TF–IDF)	Difference (test–dev)
0.10	Whole-tree coverage	0.899	0.935	+0.036
	Wrong-certificate rate	0.016	0.000	-0.016
	Certified edits / graph	0.627	1.066	+0.440
0.20	Whole-tree coverage	0.892	0.928	+0.036
	Wrong-certificate rate	0.003	0.000	-0.003
	Certified edits / graph	0.799	1.554	+0.755
0.30	Whole-tree coverage	0.901	0.915	+0.014
	Wrong-certificate rate	0.004	0.001	-0.003
	Certified edits / graph	1.110	2.223	+1.113

Earlier invalidated evaluation (run ID v2). This construction used 50/20-node units throughout, with UNESCO's minimum component at eight nodes, and yielded 219 calibration and 220 test graphs from raw pools of 120/12/120/120/14/53. All 20 evaluation cells and seven result files completed; the verification parser then rejected a certificate field exceeding its 131072-byte CSV field-size limit, and the one-shot gate transitioned irreversibly to `failed_consumed`. A post-consumption parser fix cannot resurrect a receipt, so these numbers carry no confirmatory status. Table 15 preserves every reported method and rate for historical disclosure. These outputs are not used to support the main conclusions or combined with the final evaluation.

Table 15: Historical outputs from the invalidated evaluation (run ID v2), preserved for disclosure only. WT is whole-tree coverage, Wrong is any wrong certified edit, and Edits is certified edits per graph. These outputs have no validated receipt and do not support the main conclusions.

	Rate 0.05			Rate 0.10			Rate 0.20			Rate 0.30		
Method	WT	Wrong	Edits	WT	Wrong	Edits	WT	Wrong	Edits	WT	Wrong	Edits
CARTER	0.940	0.002	0.546	0.946	0.000	0.989	0.941	0.000	1.628	0.948	0.000	2.279
Pooled edge	0.368	0.459	4.926	0.297	0.513	6.592	0.223	0.461	8.287	0.128	0.295	8.126
Bonferroni child	1.000	0.000	0.268	1.000	0.000	0.574	1.000	0.000	1.091	1.000	0.000	1.660

Superseded constructions (run IDs v3 and v4). These intermediate constructions were superseded after changes to the one-shot execution guard invalidated their frozen specifications. Their histories contain no recorded run claim, the gate event required before evaluation, and end in an unlocked state without a formal retirement event. Absence of a recorded claim does not establish that no files were read. Neither construction contributes confirmatory results. The supplement's `EVALUATION_HISTORY.md` maps all run IDs to their configurations and records, including the guard changes and the failed v4 launch.

Reserved pool for future interactive evaluation. After every exclusion — the three development benchmarks, every benchmark referenced by a historical result, and both structural holdouts — the surviving concept pool contains 128 node-disjoint instances over 5,586 nodes, 120 of these instances from OEWN noun, which is too narrow for a six-source interactive protocol. The pool remains reserved: `data/benchmark/interactive_holdout_v1` records the exclusion inventory (57,019 nodes), the per-source counts, and the instance-set digest `ad38d3b4785d1c9ec0c7a0decac91e45...`. No interactive evaluation results from this pool are reported here. An evaluation addressing the limitations in Section 6 would require extending the source coverage with fresh taxonomy releases; it is outside the present study.

D THE DESIGN SPACE OF INTERACTIVE REFINEMENT

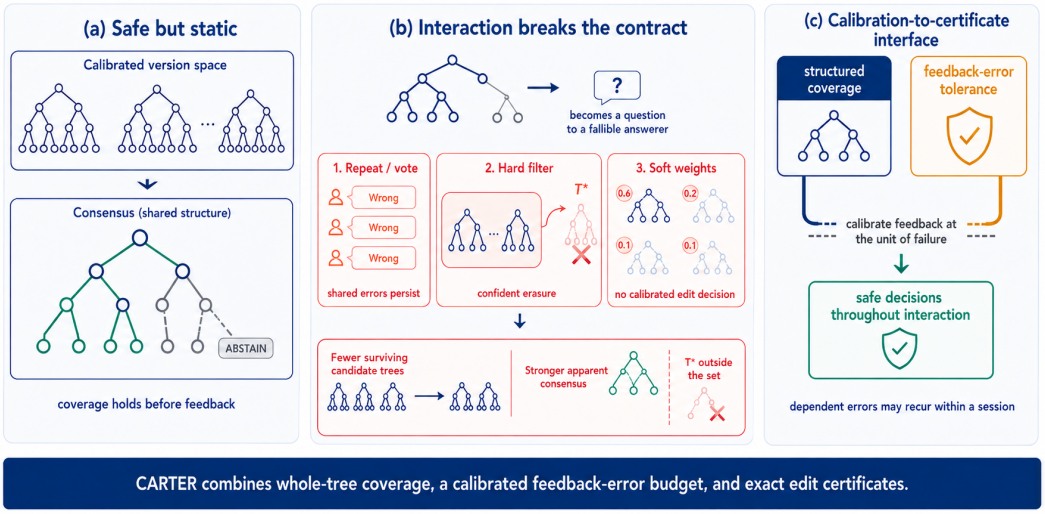


Figure 5: Design requirements for interactive structured refinement. Before feedback, a covered family of trees supports certificates for statements shared by every tree. In the illustrated workflows, repeated answers can share errors, hard filtering can erase a covered target, and answer weights alone do not specify a calibrated edit certificate. CARTER combines whole-tree coverage with a feedback-error budget calibrated over complete sessions; Figure 1 illustrates the resulting certificate interface.

E NEW BINARY-QUERY INTEGRATION DIAGNOSTICS

Design and information boundary. E1–E3 are new development-only experiments, not extensions of the sealed holdout or replacements for the canonical feedback study. The frozen plan, code hashes, input graphs, transcripts, per-prefix certificates and complete case status are released under `solution_revision/`. The input generator reads no historical dataset. E2 uses five nodes and three candidate parents per non-root child. A backbone guarantees a nonempty candidate family; the target and observed tree are independent uniform draws from its enumerated trees. Here $C = D_0$, all candidate scores and the threshold are zero, and initial target inclusion holds by construction. Thus E2 isolates feedback and computation conditional on initial inclusion; it does not evaluate candidate retrieval or structural calibration.

Policy and answerer. Each of Berlekamp one-step minimax, uniform-version entropy, and frozen random uses $b_{\mathrm{qry}} = 1$ and a four-question horizon. Entropy minimizes the exact absolute imbalance of uniform tree counts; Berlekamp minimizes the maximum successor weight including error layers and remaining horizon; both break ties lexically. Random draws legal edges with replacement. A query (u, v) maps yes to $\{u\}$ and no to the other candidate parents of v; both sets are nonempty. The utility-state fallback is the frozen random sequence if that state becomes empty. Abstention returns all candidate parents, contributes zero loss and adds no binary answer. The certificate budget never influences this query state, stopping or fallback.

The programmed iid error probability is 0.10; in the persistent regime, a child-specific error flag with that probability is shared for the whole session. A separate 0.05 abstention draw is used. Across policies, potential answers are coupled by the same graph, regime, edge and that edge's occurrence index, rather than just a common seed or round index. Different policies can therefore ask different questions. Within a policy, all certificate arms receive exactly the same transcript. Calibration uses 59 independent generated graphs and evaluation uses 60 graphs from a disjoint seed stream. Each policy and error regime receives its own complete-session calibration at $\delta_h = 0.05$. The sample sizes support a bounded integration diagnostic and a nontrivial order statistic, not precise rare-error

estimation. Repeated regimes and policies on one graph are paired observations, not additional independent graphs.

Metrics and uncertainty. Final correct edits count mandatory additions in the target and impossible deletions outside it, always relative to the original observed tree. Retained means target inclusion at every prefix; Wrong is any wrong edit at any prefix; Silent is target loss with no wrong edit over the complete session. The release separately records nonconflict silent prefixes, first target-loss and first wrong-edit rounds, exact version sizes, every edge label, and all opposing witnesses. Every prefix is checked by independent enumeration. Intervals for binary graph rates are descriptive Wilson 95% intervals. For the four prespecified primary utility contrasts, paired percentile bootstrap intervals use 10,000 graph resamples; the displayed 98.75% intervals apply a Bonferroni family-size-four adjustment. They are development analyses conditional on the realized calibration streams, not confirmatory tests or equivalence claims.

Table 16: New development-only iid simulation: 60 paired evaluation graphs per policy. Rates are reported as counts/60. Calibrated and CRC-state rows share the same selected budget and tie exactly.

Policy	Arm	b	Correct edits	Retained	Wrong	Silent	Conflict
Berlekamp	No feedback	0	0.000	60/60	0/60	0/60	0/60
Berlekamp	Hard filter	0	1.900	38/60	7/60	15/60	10/60
Berlekamp	Calibrated	1	1.000	56/60	1/60	3/60	0/60
Berlekamp	CRC-state	1	1.000	56/60	1/60	3/60	0/60
Entropy	No feedback	0	0.000	60/60	0/60	0/60	0/60
Entropy	Hard filter	0	1.083	39/60	5/60	16/60	18/60
Entropy	Calibrated	2	0.250	60/60	0/60	0/60	0/60
Entropy	CRC-state	2	0.250	60/60	0/60	0/60	0/60
Random	No feedback	0	0.000	60/60	0/60	0/60	0/60
Random	Hard filter	0	2.217	40/60	8/60	12/60	6/60
Random	Calibrated	2	0.083	60/60	0/60	0/60	0/60
Random	CRC-state	2	0.083	60/60	0/60	0/60	0/60

Table 17: New development-only persistent-child simulation: 60 paired evaluation graphs per policy. Rates are reported as counts/60. Calibrated and CRC-state rows share the same selected budget and tie exactly.

Policy	Arm	b	Correct edits	Retained	Wrong	Silent	Conflict
Berlekamp	No feedback	0	0.000	60/60	0/60	0/60	0/60
Berlekamp	Hard filter	0	2.550	42/60	11/60	7/60	1/60
Berlekamp	Calibrated	1	0.883	50/60	5/60	5/60	0/60
Berlekamp	CRC-state	1	0.883	50/60	5/60	5/60	0/60
Entropy	No feedback	0	0.000	60/60	0/60	0/60	0/60
Entropy	Hard filter	0	1.650	43/60	10/60	7/60	0/60
Entropy	Calibrated	2	0.217	57/60	1/60	2/60	0/60
Entropy	CRC-state	2	0.217	57/60	1/60	2/60	0/60
Random	No feedback	0	0.000	60/60	0/60	0/60	0/60
Random	Hard filter	0	2.117	48/60	8/60	4/60	3/60
Random	Calibrated	2	0.133	59/60	0/60	1/60	0/60
Random	CRC-state	2	0.133	59/60	0/60	1/60	0/60

Table 18: New E2 primary pipeline contrasts, Berlekamp minus the comparator: mean final correct edits and paired graph bootstrap intervals. Budgets and observed risks differ across policies, so these are not same-risk advantages.

Regime	Comparator	Difference	95% CI	Adjusted 98.75% CI
iid	Entropy	0.750	[0.483, 1.033]	[0.417, 1.117]
iid	Random	0.917	[0.617, 1.217]	[0.550, 1.317]
Persistent	Entropy	0.667	[0.400, 0.933]	[0.350, 1.000]
Persistent	Random	0.750	[0.500, 1.000]	[0.433, 1.083]

Observed calibration shortfall. The realized Berlekamp calibration chooses budget one in both regimes; the other policies choose two. Its persistent-error evaluation has 50/60 target-retained sessions and 5/60 sessions with a wrong edit. The respective Wilson intervals are [0.720, 0.907] and [0.036, 0.181]. This is an empirical safety shortfall for this fitted pipeline. The guarantee in Theorem 1 is marginal over calibration and evaluation draws, not conditional coverage for this selected budget; these observations neither establish conditional 95% safety nor alone disprove the marginal theorem. We keep the frozen budgets and all outcomes. The new loop demonstrates implementation and a safety–utility tradeoff, not superior reliability or calibrated practical advantage over entropy or CRC.

Post-hoc common-budget diagnostic. Because the primary contrasts mix query strategy and calibration-budget effects, we enumerated the saved evaluation transcripts at every fixed budget 0–4, without generating new answers or selecting a new budget. This analysis was specified after viewing E2 and remains post-hoc development evidence. The complete grid below shows that the utility ordering is not stable across budgets. In particular, entropy has more correct edits than Berlekamp at each of budgets one through three in both regimes. No fixed-budget row is assigned a newly calibrated safety guarantee.

Table 19: Post-hoc common-budget diagnostic on all saved E2 transcripts. Each cell reports the mean number of final correct edits; $n = 60$ graphs per regime and policy.

	iid			Persistent child		
b	Berlekamp	Entropy	Random	Berlekamp	Entropy	Random
0	1.900	1.083	2.217	2.550	1.650	2.117
1	1.000	1.350	0.667	0.883	1.283	0.633
2	0.033	0.250	0.083	0.033	0.217	0.133
3	0.000	0.083	0.000	0.000	0.200	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000

Exact backends and measured costs. E1 uses $n \in \{5, 8, 12\}$, $k \in \{2, 3\}$ candidate parents per child, $b_{\text{qry}} \in \{0, 1, 2\}$, two seeds, and three timing repetitions per arm. Both implementations return identical complete layer coefficients, all candidate-query values at horizons 1, 3, and 4, and lexical selections on all 36 paired inputs; the five-node cases also match enumeration. The median, across inputs, of the ratio of the reference median compute-plus-serialization time to the shared-compiler median time is 2.407. This is a computational comparison of equal outputs, not an accuracy or query-utility improvement.

E3 compares the original certificate solver with the existing optimized solver using one or two worker processes. Full dataclass equality includes minimum cost, canonical baseline tree, all edge sets, restricted optima, canonical witnesses, conflict and logical oracle calls. Twelve inputs have completed reference outputs and pass this entire comparison. For two 100-node/$k = 5$ inputs, only the optimized arms complete, so their optimized outputs are not reference-validated and support no upgrade claim. All six 500-node jobs time out. Each job has a 45-second limit including imports, three repetitions on regular inputs or one on 500-node inputs, and process teardown; a job timeout is not a measured solver completion time.

Table 20: E3 full-certificate compute plus serialization, median of per-input medians in seconds. Each size/density has two input seeds; IQRs, individual times and sampled process-tree RSS are in the JSON release. T/O means the complete job did not finish; no censored speed ratio is computed. All costs include the solves that generate opposing witness trees.

n	k	Reference	Optimized (1)	Optimized (2)	Reference parity
5	3	0.007	0.004	2.610	2/2 pass
20	3	0.076	0.018	2.624	2/2 pass
50	3	1.820	0.109	2.933	2/2 pass
100	3	8.491	1.279	3.590	2/2 pass
20	5	0.180	0.051	2.722	2/2 pass
50	5	4.269	0.301	3.084	2/2 pass
100	5	T/O	1.905	4.160	Unavailable
500	5	T/O	T/O	T/O	Unavailable

The optimized implementation reuses graph preprocessing and an exact unique-incoming shortcut, falling back to the reference branching routine. Restriction time includes generating alternative trees; the separately reported witness-materialization field is only wrapper construction and must not be interpreted as the complete cost of obtaining witnesses. Two-process runs include pool startup and can be slower on small inputs. Fresh-worker launch, imports and teardown are included in the outer job timer, not separated as pure process-startup cost. RSS is sampled over the worker process tree every 20 ms; these samples do not provide an exact memory bound. E1 and early E3 jobs overlapped the E2 correctness run; these are ambient-load desktop measurements, not isolated-machine latency claims. Hardware and runtime versions are recorded from the executing machine in the run manifest. The old 152-second reference observation and older 3.58-second optimized median remain historical, unmatched measurements. Neither those observations nor this new grid supports full-MeSH or 500-node exact-query interaction.

Executable demonstration and evidence separation. The loopback-only demonstration has a verified replay of Figure 1 and a manual-answer exploration mode. Replay is labeled as a constructed mechanism, with no population coverage claim. Exploration is explicitly uncalibrated: changing answers or policy does not inherit the paper's calibrated guarantee. Every displayed edit, conflict and witness is returned by the real certificate solver. The demo is a workflow artifact, not human usability evidence or an LLM memory experiment. The existing executable output-safe/state-lost counterexample is also rechecked; it establishes an event distinction, not superiority over general CRC or an empirical occurrence rate.

F Independent repeated-calibration diagnostic

Question and protocol. This new development-only programmed simulation was designed after observing E2's single-fit shortfall. Its protocol, seed schedule, sample sizes and analysis were fixed before new outcomes were generated; it neither replaces E2 nor accesses the sealed holdout. We repeat the complete calibration and evaluation 100 times, with 59 calibration graphs and 60 independent evaluation graphs per repetition, retaining E2's five-node, three-candidate-per-child generator, horizon four, $b_{\mathrm{qry}} = 1$, $\delta_h = 0.05$, error probability 0.10 and abstention probability 0.05. Candidate inclusion is imposed and $C = D_0$, so this tests feedback calibration, not structural calibration. Targets and observed trees are independent uniform draws from each generated graph's valid rooted trees. Fresh hash-separated streams distinguish every outer repetition and calibration/evaluation split. Graph structures may coincide by chance; independence concerns the random draws, not unique node names or exclusion of repeated structures.

Each regime/policy has 5,900 calibration and 6,000 evaluation sessions across 100 fits. The three policies and two regimes share generated graph units and the specified potential-answer mechanism, not necessarily questions or transcripts. The highest independent analysis unit is the outer repetition; the 71,400 total policy/regime session records are not 71,400 independent calibration fits. The fixed 100-repeat design diagnoses between-fit variability and coarse budget-selection frequencies, not small equivalence margins or human performance. All repetitions and fixed budgets zero through four are retained, with no outcome-based exclusions or expansions.

Implementation and checks. An independent finite-tree enumerator implements the same queries for this small diagnostic. Before outcome generation, it reproduced all 714 saved E2 development trajectories, including 2,856 query choices and answers. The new run's 71,400 session scores, 600 budgets and 900,000 prefix states across the full budget grid pass an independent scalar/set replay that calls neither the production solver nor its metric implementation. This enumeration is not a scalable solver upgrade or a runtime benchmark. The separately computed binary CRC criterion selects the same budget as whole-session calibration in every fit, hence its use with the same transcript and exact certificate map gives identical states and edits, not an additional empirical advantage for either name.

Estimation. Rates and final correct-edit counts are averaged over 60 evaluation sessions within each fit and then over fits. Central two-sided percentile intervals use 10,000 outer bootstrap resamples, paired across policies. Four prespecified utility contrasts compare the calibrated Berlekamp pipeline against entropy and frozen random separately in each regime; their Bonferroni-level intervals use 98.75% marginal bootstrap levels. These are approximate bootstrap intervals, not exact finite-sample coverage statements. With separately calibrated budgets, these are pipeline contrasts, not isolated query-policy or matched-risk comparisons. Secondary intervals are descriptive 95% intervals. No non-significant contrast is interpreted as equivalence.

Table 21: New repeated-calibration results: 100 outer fits, 6,000 evaluation sessions per row. Brackets show descriptive outer-cluster 95% intervals. "Correct" counts final correct additions plus deletions, not all certified edits. The budget column gives counts for $b = 1/2/3/4$.

Regime	Policy	Budgets	Retained	Wrong	Correct
iid	Berlekamp	49/51/0/0	0.976 [0.970, 0.981]	0.009 [0.007, 0.012]	0.498 [0.408, 0.589]
iid	Entropy	42/58/0/0	0.979 [0.973, 0.985]	0.010 [0.007, 0.013]	0.659 [0.540, 0.777]
iid	Random	50/50/0/0	0.979 [0.973, 0.984]	0.008 [0.006, 0.010]	0.316 [0.265, 0.368]
Persistent	Berlekamp	3/79/18/0	0.976 [0.969, 0.981]	0.006 [0.004, 0.009]	0.076 [0.050, 0.108]
Persistent	Entropy	0/75/6/19	0.979 [0.975, 0.983]	0.006 [0.004, 0.008]	0.044 [0.037, 0.051]
Persistent	Random	8/69/22/1	0.971 [0.965, 0.977]	0.009 [0.005, 0.013]	0.078 [0.056, 0.103]

Table 22: All four prespecified paired pipeline utility contrasts. Difference is Berlekamp minus the comparator in correct edits per graph. The last column uses Bonferroni-level 98.75% outer-bootstrap intervals.

Regime	Comparator	Difference	95% interval	98.75% interval
iid	Entropy	-0.1610	[-0.2950, -0.0297]	[-0.3329, 0.0055]
iid	Random	0.1813	[0.0747, 0.2895]	[0.0443, 0.3211]
Persistent	Entropy	0.0313	[0.0057, 0.0648]	[-0.0001, 0.0767]
Persistent	Random	-0.0020	[-0.0375, 0.0373]	[-0.0460, 0.0500]

Interpretation. The original persistent-error budget-one result is not replaced: this budget is selected in 3 of the 100 new Berlekamp fits, while 79 select two and 18 select three. Marginal retention in the new simulation therefore coexists with much lower useful-edit counts than the original single-fit result. Across the four multiplicity-adjusted utility comparisons, only iid Berlekamp versus random has an interval entirely above zero; neither Berlekamp–entropy comparison establishes an advantage at the adjusted level. This is evidence about the specified programmed pipelines, not conditional coverage for every fitted budget, human effectiveness, or a general query-policy ranking. The raw all-budget metrics also preserve target-loss rounds, wrong-edit rounds, conflict, silent loss and conditional correct-edit numerators and denominators for further descriptive audit.

Artifacts. `development/repeated_calibration/PROTOCOL.json` defines the frozen design. The `repeated_calibration/` directory contains raw compressed session and per-prefix metric records, fit summaries, calibration scores, run and code hashes, independent replay and aggregate result files. Results in this section are generated from that registry; the anonymous release contains the runner, independent auditor and result builder.

G POST-HOC RELEASED-OUTPUT SENSITIVITY

The frozen primary estimand in Table 1 gives the six source strata equal weight, including the small ANZSRC FoR slice. To show whether that slice drives the descriptive pattern, we reaggregated the released final-v5 per-instance outputs after omitting ANZSRC FoR. This is a post-hoc released-output sensitivity: it does not reopen the consumed holdout, rerun the scorer, change the six-source primary estimand, or add intervals or p-values. We retain the original aggregation order (five frozen seeds within graph, graphs within source, then equally weighted source means). The released JSON records all six per-source means. For utility, the table reports *canonical-correct* certified edits rather than all certified edits, so a gain does not count an incorrect certificate.

Table 23: Post-hoc five-source sensitivity excluding ANZSRC FoR. Coverage and wrong are CARTER whole-tree coverage and graph-level any-wrong-certificate rates. Gains are CARTER canonical-correct certified edits per graph minus the indicated conservative baseline. These released-output summaries are descriptive and do not replace the six-source primary results.

Corruption	Coverage ↑	Wrong ↓	Gain vs. Bonf. ↑	Gain vs. LTT ↑
0.05	0.915	0.001	0.418	0.418
0.10	0.922	0.000	0.590	0.590
0.20	0.914	0.000	0.719	0.719
0.30	0.897	0.001	0.949	0.949

The five-source coverage spans 0.897–0.922, the observed wrong rate remains at most 0.001, and the correct-edit gain remains positive at every displayed corruption rate. This is consistency evidence for the reported fixed-source summary, not conditional coverage, a new confirmatory analysis, or unseen-domain generalization. Separately, on the saved canonical development-only simulated-feedback sessions, the calibrated arm has 3.432 final canonical-correct edits per session versus 1.345 for no feedback, a paired descriptive increment of 2.088. This correct-edit summary is distinct from Table 2, whose displayed increment counts all final certified edits; it remains neither a human/LLM result nor a new coverage analysis.