DEPENDENCY HIERARCHY GOVERNS CURVATURE GAINS IN STRUCTURED LATENT REPAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

Dependency hierarchy conditions when curved geometry improves structured latent repair: branching traces provide metric expansion for selective updates, whereas flat connectivity offers little structure to exploit. We test this prediction with CURVREPAIR through a confound-closed audit that fixes graph source, support retrieval, locator, pullback, update norm, trigger, and selection rule while changing only the latent metric. On the strict released-only `ProofWriter-OWA`+`GSM-Hard`+`MATH-500` endpoint (n=1,242), CURVREPAIR gains $+1.55$ percentage points over its edit-rule-matched Euclidean control (95% CI $[+1.29, +1.81]$, paired-bootstrap p=6.8×10^{-6}) while satisfying the prespecified 0.15 pp harm tolerance. Matched rewiring separates the structural regimes: depth-preserving rewiring retains 85.2% of the whole-set gain, hierarchy destruction removes 68.4%, and random edges leave an 11.6% residual whose interval includes zero. The effect replicates across the complete 23-task `BIG-Bench-Hard` release (n=6,511; $+1.68$ pp), the complete MMLU-Pro release (n=12,032; +1.46 pp, 95% CI [+1.19,+1.73]), and a fixed-anchor repair workflow (+0.94 pp). At equal end-to-end token budgets, CURVREPAIR leads Self-Consistency-Rewrite, Reflexion-Plus, and Constrained-Regeneration, with all three paired intervals above zero. Confound closure and structural intervention therefore turn hierarchy into an actionable criterion for choosing curved geometry in latent repair.

1 INTRODUCTION

Local latent repair is intrinsically structural. Correcting one reasoning step should influence its dependencies while leaving unrelated branches stable, so the shape of the dependency graph constrains which latent metric can express a selective update. Hyperbolic spaces offer exponential volume for separating sibling branches without pushing descendants away from their ancestors (Nickel & Kiela, 2017; Ganea et al., 2018; Chami et al., 2019). This geometry suggests a concrete design principle: curvature should yield its largest return on deep, branching traces and lose that return as hierarchy is removed.

Testing that principle requires more than comparing a hyperbolic package with a Euclidean baseline. Graph supervision, support retrieval, localization, centroid construction, update normalization, and decoding can each change the answer. We therefore define *confound-closed geometry auditing*: the graph source, support bank, locator, pullback, edit rule, parameter budget, update-strength sweep, trigger, and collateral constraint are frozen across arms. The resulting paired whole-set accuracy difference is an identifiable metric-choice estimand over every evaluation example.

CURVREPAIR instantiates the curved arm. It maps an intermediate residual state to a Poincaré ball, moves the selected step toward a train-only Fréchet target, and returns the normalized tangent displacement through the projection Jacobian. Edit-Rule-Matched Euclidean ‡ executes the same graph losses, support selection, log–exp barycentric solver, Jacobian-transpose pullback, trigger, and residual-norm rule in Euclidean space. This counterfactual design isolates the consequence of geometry inside a fixed intervention.

The strict decision endpoint combines trace-only graphs and benchmark-excluded retrieval over all 1,242 held-out examples from the released benchmarks `ProofWriter-OWA`, `GSM-Hard`, and `MATH-500`, four instruction-tuned backbones, five seeds, and a frozen `DeBERTa-v3-base` locator.

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.

CURVREPAIR improves whole-set accuracy over ‡ by $+1.55$ pp (95% CI $[+1.29, +1.81]$, paired-bootstrap $p{=}6.8\times10^{-6}$) and satisfies the prespecified 0.15 pp initially-correct harm tolerance. Each released domain and all four backbones retain the positive ordering; a locator-free entropy trigger yields a $+1.27$ pp matched gain.

Matched graph interventions reveal what drives the result. Rewiring within depth bands retains 85.2% of the strict whole-set accuracy advantage, hierarchy-destroying rewiring removes 68.4%, and random connectivity leaves an 11.6% residual whose confidence interval crosses zero. Separate depth-stratified, controlled-structure, direct-recovery, and edit-locality analyses reproduce this ordering under their respective estimands. Hierarchy is therefore a measurable moderator of curvature's value.

The finding extends across benchmark families. The matched advantage is $+1.68$ pp on the complete 23-task BBH release ($n{=}6{,}511$) and +1.46 pp on all 12,032 questions from MMLU-Pro (Wang et al., 2024); 11/14 MMLU-Pro disciplines are positive.

It also transfers across inference and repair construction. At equal end-to-end token budgets, CURVREPAIR leads SC-RW, Reflexion-Plus, and Constrained-Regeneration with resolved paired intervals. TangentRoute, a fixed-anchor correction-direction workflow, preserves a +0.94 pp curved-over-Euclidean gain. Together these results make hierarchy a practical model-selection variable for local latent repair.

Our contributions are:

- *A structural finding.* Controlled rewiring, depth stratification, and a fixed-structure task show that dependency hierarchy governs the accuracy and locality return from curvature.
- *An identifiable metric-choice test.* Confound-closed geometry auditing fixes the major gain-producing levers around a local latent edit, making the curved-versus-Euclidean whole-set gap a falsifiable estimand.
- *A replicated decision boundary.* The hierarchy-conditioned advantage persists across three core domains, four backbones, BBH, MMLU-Pro, locator-free triggering, transferred retrieval, exact-budget non-local inference, and a fixed-anchor workflow.

This combination of structural intervention and confound closure identifies when curvature changes repair quality and supplies a reusable standard for geometry-sensitive editing. Section 2 defines the estimand, Section 3 constructs the matched geometry pair and structural hypothesis, and Section 4 evaluates attribution, transfer, and hierarchy dependence.

2 CONFOUND-CLOSED GEOMETRY AUDITING

Estimand. Confound-closed geometry auditing freezes the edited workflow, graph source, support bank, locator, pullback, edit rule, optimization budget, and trigger, then varies only latent geometry between treatment and decisive control. The estimand is the paired whole-set task-accuracy difference at a prespecified collateral operating point. For a reasoning trace, let language model f supply hidden states $\mathbf{h}_1, \ldots, \mathbf{h}_T$ and let $\mathcal{G} = (V, E)$ relate its reasoning steps; the curved arm learns $\pi_\theta : \mathbb{R}^d \to \mathbb{H}^n$ in the Poincaré ball $\mathbb{H}^n = \{\mathbf{z} \in \mathbb{R}^n : \|\mathbf{z}\| < 1\}$:

$$d_{\mathbb{H}}(\mathbf{z}_i, \mathbf{z}_j) = \operatorname{arcosh}\left(1 + 2\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{(1 - \|\mathbf{z}_i\|^2)(1 - \|\mathbf{z}_j\|^2)}\right). \tag{1}$$

Relation graphs. Clean-graph conditions construct $\mathcal{G}$ from four task-side sources: (1) `tree-sitter-0.20.8` AST def–use dependencies for code; (2) released `ProofWriter-OWA` proof trees; (3) CPython 3.11 `sys.settrace` execution dependencies for arithmetic; and (4) released `prm800k` decomposition trees for `MATH-500`. These graphs use neither model attention nor generated answers. The strict condition instead uses the answer-masked trace-only construction specified in Section 4 and Appendix A.5.

Support set and anti-leakage rule. Each editable step t receives $K{=}16$ retrieved embeddings of correct steps from `train`; for `MATH-500`, the source is disjoint `prm800k` IDs. Candidates match benchmark, backbone, coarse step index, and local graph signature, then rank by pre-edit residual

Table 1: Audit invariants for the decisive CURVREPAIR versus Edit-Rule-Matched Euclidean ‡ comparison. Geometry is the sole varied factor.

Component	Shared setting	Varied factor
Graph source	Same nodes, edges, and edge labels	—
Support bank	Same train-only K=16 candidates	—
Locator and trigger	Same step and firing decision	—
Pullback	Same Jacobian-transpose rule and norm target	Metric-trained head
Losses and weights	Same terms and $\lambda_s, \lambda_c, \lambda_g$	Metric in distances
Parameter/training budget	Matched head size and update schedule	—
Edit-strength sweep	$\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$	—
Collateral rule	Same ≤ 0.25 selection constraint	—
Latent geometry	Matched n=32 representation	Hyperbolic / Euclidean

similarity. A development-fixed degree fallback applies when fewer than 16 exact-signature matches exist. No `dev`, `locator-train`, `repair-eval`, or current-example step enters a prototype. Four transfer conditions test cross-backbone, cross-benchmark, target-benchmark-excluded, and jointly target-benchmark/backbone-excluded banks (Table 45).

Objectives and parameter updates. For the gold continuation tokens $y_{t:}$ following an edited step, $\mathcal{L}_{\text{task}} = -\sum_{k \geq t} \log p_f(y_k \mid y_{<k}, \tilde{\mathbf{h}}_t)$ is the task-native next-token loss after the pulled-back intervention. The geometric term is

$$\mathcal{L}_{\text{geo}} = \lambda_s \underbrace{\sum_{(i,j) \in E_{\text{dep}}} d_{\mathbb{H}}(\mathbf{z}_i, \mathbf{z}_j)^2}_{\text{dependency smoothness}} + \lambda_c \underbrace{\sum_{(i,j) \in E_{\text{contra}}} \max(0, \delta - d_{\mathbb{H}}(\mathbf{z}_i, \mathbf{z}_j))}_{\text{contradiction separation}} + \lambda_g \underbrace{\sum_{i \in V} \ell_{\text{kNN}}(\mathbf{z}_i, \mathcal{G})}_{\text{graph consistency}}, \quad (2)$$

where ℓ_{kNN} is binary cross-entropy between the latent k-nearest-neighbor graph (k=4) and the dependency neighborhood in $\mathcal{G}$. We fix λ_s=1.0, λ_c=0.5, λ_g=0.25, and δ=2.0 on `ProofWriter-OWA` development data and reuse them everywhere. During projection-head training, $\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{geo}}$ updates θ while f remains frozen. During joint tuning, $\mathcal{L}_{\text{task}}$ updates both θ and the lightly tuned backbone parameters, while $\mathcal{L}_{\text{geo}}$ updates θ only; this prevents graph supervision from directly changing the language-model backbone. The Euclidean arm follows the identical update partition.

Edit desideratum. On a useful editing surface, intervening at step t should change states in t's graph neighborhood much more than those beyond it. Write $N_1(t)$ for the dependency neighborhood one hop from t, and define $U(t) = \{j > t : j \notin N_1(t)\}$. For tokens aligned with step j, Δ_j denotes the mean token-level Jensen–Shannon divergence comparing next-token distributions before and after the edit. The reported measures are

$$\text{Locality}(t) = \frac{\sum_{j \in N_1(t) \cup \{t\}} \Delta_j}{\sum_{j=1}^{T} \Delta_j}, \qquad \text{Collateral}(t) = \frac{1}{|U(t)|} \sum_{j \in U(t)} \mathbf{1}\left[\hat{y}_j^{\text{edit}} \neq \hat{y}_j\right], \quad (3)$$

where $\hat{y}_j$ is the greedy decoded text of step j. Contradiction separation is the mean contradiction-pair distance divided by the mean dependency-pair distance within each geometry, making the ratio comparable across metrics. Structural measures accompany task-native end-task recovery and whole-set accuracy throughout.

3 CURVREPAIR AND THE MATCHED EUCLIDEAN CONTROL

CURVREPAIR instantiates the curved treatment; ‡ instantiates the counterfactual obtained by replacing hyperbolic operations with their Euclidean counterparts while preserving every invariant in Table 1. This pairing makes the whole-set difference directly interpretable as the residual consequence of geometry within the evaluated repair operator.

3.1 ARCHITECTURE

We implement projection head π_θ as a 3-layer MLP: 512 hidden units and an exponential map activation place its output on the Poincaré ball. Its input is residual-stream hidden state $\mathbf{h}_t$ from the

lightly tuned backbone (`Qwen-2.5-7B-Instruct` (Qwen Team, 2024)), using layer 20 among its 28 layers. The projection dimension is $n = 32$. A single choice on `dev` fixed the layer and operating dimension; the locality–latency trade-off near this setting appears in Appendix B.12.

Geometric hypothesis. Hyperbolic volume grows exponentially with radius, allowing deep descendants to remain close to parents while separating sibling branches. A correction can therefore approach its local prototype without crowding unrelated branches into the same neighborhood. The hierarchy rewiring experiment in Section 4 tests the corresponding prediction: the curved advantage should persist when depth is preserved and contract when hierarchy is destroyed.

3.2 Training and Intervention

With backbone f frozen, π_θ undergoes 10,000 training steps with `lr`$= 5\times10^{-4}$; Riemannian Adam optimizes $\mathcal{L}=\mathcal{L}_{\text{task}}+\mathcal{L}_{\text{geo}}$, with the radius clipped to 0.95. Joint tuning then runs for 5,000 steps with backbone `lr`$= 2\times10^{-5}$. For each backbone, we cache training-pool support banks once and reuse these banks during inference. At the located error step t, we target the Fréchet centroid of its support bank S_t:

$$\bar{\mathbf{z}}_t = \arg\min_{\mathbf{y}\in\mathbb{H}^n} \sum_{i\in S_t} d_{\mathbb{H}}(\mathbf{y}, \mathbf{z}_i)^2, \qquad \tilde{\mathbf{z}}_t = \operatorname{Exp}_{\mathbf{z}_t}\left(\alpha \operatorname{Log}_{\mathbf{z}_t}(\bar{\mathbf{z}}_t)\right), \tag{4}$$

where α is the intervention strength. Let $\mathbf{v}_t = \operatorname{Log}_{\mathbf{z}_t}(\tilde{\mathbf{z}}_t)/\|\operatorname{Log}_{\mathbf{z}_t}(\tilde{\mathbf{z}}_t)\|_{\mathbf{z}_t}$ be the unit tangent direction. We pull it into the residual stream as $\delta\mathbf{h}_t = \eta\alpha J_{\pi_\theta}(\mathbf{h}_t)^\top \mathbf{v}_t$, normalize the resulting residual displacement to the development-set-fixed target norm, and set $\tilde{\mathbf{h}}_t = \mathbf{h}_t + \delta\mathbf{h}_t$ before resuming decoding. The Jacobian transpose is evaluated at the same pre-edit state in both arms.

Edit-Rule-Matched Euclidean ‡. The control uses an n=32 Euclidean projection with the same architecture, parameter count, initialization schedule, optimizer budget, and five-step barycentric solver. It substitutes Euclidean distance and Euclidean log/exp maps while keeping S_t, α, the Jacobian-transpose construction, and residual-norm target fixed. The latent metric is the sole ex ante intervention; each arm's learned head and tangent direction are metric-induced outcomes. Appendix B.2 gives the full implementation grid.

3.3 Hierarchy Hypothesis and Matched Structural Intervention

Before strict-setting evaluation, we fix graph depth, branching, and a per-benchmark Gromov δ-hyperbolicity estimate $\bar{\delta}(\mathcal{G})$ on training or disjoint proxy IDs. Across six natural benchmark families, deeper branching with lower $\bar{\delta}$ tracks a larger CurvRepair$-$‡ gap. The controlled unbranched-chain condition further separates path length from branching hierarchy. We test the intervention-level prediction by rewiring the same released graphs while holding size, edge-type marginals, support, optimization, and evaluation fixed. Appendix E.2 reports the fixed diagnostics and controlled structure task; Appendix E.7 reports the matched intervention.

4 Experiments

This section evaluates confound-closed geometry in decision order: the strict released-only endpoint, exact-budget non-local competitors, external-family replications, and interventions that locate the structural source of the effect. Table 2 is the decision criterion; subsequent results measure practical competitiveness, breadth, workflow transfer, and hierarchy dependence.

4.1 Setup

Benchmarks and strict held-out discipline. (1) `CodeContests-Debug` (1,247 used / 1,247 released): our debugging slice of CodeContests (Li et al., 2022), with Python-3 WA/accepted examples and AST graphs from `tree-sitter-0.20.8`; (2) `ProofWriter-OWA` (2,120 / 2,120): the official dev set at `OWA depth-5`, including its proof-tree annotations (Tafjord et al., 2021); (3) `GSM-Hard` (1,593 / 1,621): the harder GSM8K-style split of Gao et al. (2023), itself derived from GSM8K (Cobbe et al., 2021), with 28 deterministic exclusions under `sys.settrace` CPython-3.11 execution; (4) `MATH-500` (500 / 500): the fully standard evaluation slice `MATH-500` (Light-

man et al., 2024), released as part of the `prm800k` line and used without any paper-defined subslicing. We apply a one-time SHA-256 partition of upstream IDs to `CodeContests-Debug`, `ProofWriter-OWA`, and `GSM-Hard`. The four resulting disjoint pools allocate 60% to `train`, 10% to `dev`, 10% to `locator-train`, and 20% to `repair-eval`. Because the released `MATH-500` evaluation slice is already fixed, it follows an exception: `repair-eval` retains all 500 items as held-out data. Its `train`/`dev`/`locator-train` examples instead come from disjoint `prm800k` IDs, with zero overlap in hashed IDs against `MATH-500`. Appendix A.8 details this split rule and reports 0 pairwise overlap; it also lists the `repair-eval` domain counts, 249/424/318/500, totaling $1{,}491$ held-out examples. The debugging slice is study-defined; ProofWriter and GSM-Hard are partitioned as described above, while MATH-500 uses the full released evaluation slice.

Metrics and inference. We report task accuracy, edit locality and collateral damage from Eq. 3, contradiction separation, direct recovery, and $1/2$-hop confinement. Repair-only tables report the post-edit success rate over candidate failures; whole-set tables report recovery on the initially incorrect subset. Subscripted $\pm$ values denote bootstrap standard errors; bracketed ranges and plotted bands are 95% percentile-bootstrap intervals from $1{,}000$ paired item resamples within benchmark/backbone cells. Reported p-values use a separate two-sided paired-bootstrap test with 10^8 null-centered resamples, or exact enumeration when the discordant-pair count permits it. The task-native criterion for `CodeC-Dbg` is unit-test pass; `ProofWriter-OWA` requires proof validity, while `GSM-Hard`/`MATH-500` use exact match.

Matched-budget protocol. For each method, the sweep covers $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. We report direct recovery at the strongest operating point subject to pooled collateral no greater than 0.25. When several points meet this budget, selection favors the highest direct recovery. Thus, simply rewriting more of a trace cannot win under this protocol. Every backbone and repair regime follows this budget rule; in the strict released-only comparison, an additional latency sweep over three budgets verifies the ordering at practical operating points.

Whole-set trigger and harm metric. The primary whole-set endpoint contains all $1{,}242$ examples in the released-only pool; broader appendix diagnostics additionally include `CodeC-Dbg` for $1{,}491$ held-out examples. The shared frozen `DeBERTa-v3-base` locator fires when predicted error exceeds the development-fixed τ=0.42, producing an 18.4% edit rate on the strict endpoint; a verifier-free operating point uses τ=0.36 at the same approximate rate. The locator-free test instead fires at $\max_t \bar{H}_t \geq 1.38$ nats and edits $\arg\max_t \bar{H}_t$, yielding a $\approx 14\%$ rate. We report whole-set accuracy, recovery on initially incorrect traces, accuracy change on initially correct traces (tolerance ≤ 0.15 pp), edit rate, and end-to-end and per-call latency on `Qwen-2.5-7B-Instruct` with `A100-80GB`, sequence length $2{,}048$, and batch 32. Within each protocol, all methods receive identical trigger decisions.

Baselines. All local controls share optimizer family, training schedule, pullback, intervention sweep, and trigger. Euclidean Graph-Aware Editor † matches graph supervision; Edit-Rule-Matched Euclidean ‡ additionally matches the tangent-space barycentric rule and is the geometry control. Residual Steering •, Step-local Regenerate ⋄, and Error-aware Prompted Rewrite ⋄⋄ measure lighter local alternatives. Non-local comparisons use our budget-matched instantiations of Constrained-Regeneration ⧫ (beam-4 constrained suffix rewriting), Reflexion-Plus ⧫⧫ (two-round verbal self-critique (Shinn et al., 2023)), and Self-Consistency-Rewrite (eight sampled suffixes with majority vote (Wang et al., 2023)). CR and RX+ span $+3\%$, $+8\%$, and $+15\%$ per-call latency; token-matched rows differ by at most 0.8% in end-to-end tokens. Manifold and graph-corruption controls isolate geometry and structure.

Trace-only universal graph. For the three released benchmarks, the strict graph uses the segmented pre-answer trace, public problem text, deterministic adjacency/coreference/numeric-reuse edges, and frozen `cross-encoder/nli-deberta-v3-base` contradiction scores. The final-answer node is located by a lexical prefix, and its generated answer span is masked before edge extraction and pair classification. No AST, proof tree, execution trace, worked solution, gold answer, reference solution, or LLM teacher enters this condition. The pool contains 424+318+500=1,242 examples. For the 40 traces whose segmentation fails, every arm executes the same no-edit fallback; those IDs remain in whole-set accuracy, bootstrap, and transition denominators.

Repair regimes, weak graphs, and compute. We evaluate three repair regimes: (a) *oracle* uses a step flagged by the verifier; (b) *predicted* uses the shared `DeBERTa-v3-base` locator, frozen after training only on `locator-train` (step-level F1 values of $0.764/0.802/0.781/0.756$); regime (c) *verifier-free predicted* covers all $4{,}619$ held-out traces that fail the direct criterion, without invoking a verifier. Non-privileged graphs are either a 24M-parameter `DeBERTa-v3-xsmall` inducer trained solely from token co-occurrence and shared-identifier heuristics or the uniform trace-only construction in the preceding paragraph. The primary backbone is `Qwen-2.5-7B-Instruct`; transfer uses `Llama-3.1-8B-Instruct` (Dubey et al., 2024), `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023), and `Qwen-2.5-3B-Instruct`. Every experiment uses 8×`NVIDIA A100 80GB`, driver `535.104.12`, CUDA 12.4, PyTorch 2.3.1, `transformers` 4.42.3, and `geoopt` 0.5.1. We run seeds $\{3, 7, 13, 23, 42\}$ and $1{,}000$ paired-bootstrap resamples. Reported experiments consume 942 A100-hours (994 including pilots), including 45 hours for BBH. Appendix A.8 gives prompts, decoding, segmentation, and compute accounting.

4.2 Results

Decision criterion and evidence hierarchy. The primary endpoint is the paired whole-set accuracy gap between CurvRepair and Edit-Rule-Matched Euclidean ‡ in Table 2's strictest row, with identical trigger, update, and 0.25 collateral-selection rules. CurvRepair must additionally satisfy the prespecified 0.15 pp initially-correct harm criterion. Exact-budget non-local methods quantify practical competitiveness; BBH and MMLU-Pro test external breadth; locator-free and transfer settings test robustness; matched rewiring tests the hierarchy prediction.

RQ1: What curvature advantage remains for hierarchy to explain? With trace-only graphs, CurvRepair gains $+1.81$ pp; with benchmark-excluded retrieval, it gains $+2.29$ pp. Combining both restrictions yields the strict $+1.55$ pp effect (95% CI $[+1.29, +1.81]$, paired-bootstrap $p{=}6.8\times10^{-6}$), with $+2.04/+1.58/+1.12$ pp on `ProofWriter-OWA`/`GSM-Hard`/`MATH-500`. All three released domains and all four backbones preserve the positive ordering, with every evaluation example retained in the estimand. Initially-correct harm is -0.10 pp, inside the prespecified tolerance, while recovery on initially incorrect examples rises by 9.41 pp.

Table 2: Strict released-only whole-set utility on `ProofWriter-OWA`+`GSM-Hard`+`MATH-500` (n=1,242; four backbones, five seeds). The strictest row combines trace-only graphs and benchmark-excluded retrieval. Gaps are paired against ‡; every curved row satisfies the prespecified 0.15 pp harm criterion.

Condition	CurvRepair Acc. (%)	Δ vs. ‡	‡ harm	Curved harm	Recovery (%)
Trace-only graph	$58.76_{\pm0.19}$	$+1.81_{\pm0.13}$	–	$-0.11_{\pm0.06}$	$+9.73_{\pm0.28}$
Benchmark-excluded bank	$59.24_{\pm0.19}$	$+2.29_{\pm0.13}$	–	$-0.09_{\pm0.06}$	$+10.42_{\pm0.28}$
Both (strictest)	$58.50_{\pm0.19}$	$+1.55_{\pm0.13}$	$-0.20_{\pm0.07}$	$-0.10_{\pm0.06}$	$+9.41_{\pm0.28}$

Condition	ProofWriter Δ‡	GSM-Hard Δ‡	MATH-500 Δ‡
Trace-only graph	$+2.36_{\pm0.28}$	$+1.84_{\pm0.30}$	$+1.36_{\pm0.21}$
Benchmark-excluded bank	$+2.92_{\pm0.28}$	$+2.31_{\pm0.30}$	$+1.74_{\pm0.21}$
Both (strictest)	$+2.04_{\pm0.27}$	$+1.58_{\pm0.29}$	$+1.12_{\pm0.21}$

Whole-set inference and transitions. Paired-bootstrap p-values are 6.8×10^{-6}, 1.1×10^{-6}, 7.2×10^{-4}, and 0.21 for original, depth-preserving, hierarchy-destroying, and random graphs. Across 24840 paired example–backbone–seed outcomes, Euclidean-wrong/curved-correct, curved-wrong/Euclidean-correct, both-correct, and both-wrong counts are 1480/1095/13051/9214, 1342/1014/13100/9384, 810/688/13399/9943, and 672/627/13452/10089. Hierarchy destruction removes 68.4% of the primary whole-set accuracy gap while preserving all audit invariants.

RQ2: Does hierarchy govern the curvature advantage? Depth-preserving rewiring retains 85.2% of the strict whole-set accuracy effect, hierarchy destruction leaves 31.6%, and random connectivity leaves a nonsignificant 11.6%. The paired transition ledger and bootstrap tests in Table 3 separate the preserving and destroying conditions, while the direct-recovery and edit-locality analyses in

Table 3: Cross-workflow and hierarchy interventions on the exact primary whole-set endpoint (n=1,242). Every row uses trace-only graphs, the shared locator, four backbones, five seeds, and matched update and collateral-selection budgets. Rewiring preserves node, edge, and edge-type counts.

Condition	D	B	δ	Gap [95% CI]	Retained
Original trace graph	3.84	2.11	0.34	$+1.55$ $[+1.29, +1.81]$	100%
Fixed-anchor workflow transfer					
TangentRoute workflow	3.84	2.11	0.34	$+0.94$ $[+0.67, +1.21]$	–
Graph intervention from the original trace graph					
Depth-preserving	3.79	2.07	0.37	$+1.32$ $[+1.06, +1.58]$	85.2%
Hierarchy-destroying	1.12	0.98	1.27	$+0.49$ $[+0.21, +0.77]$	31.6%
Random edges	1.05	1.04	1.31	$+0.18$ $[-0.10, +0.46]$	11.6%

Condition	Euc. acc.	Curved acc.	Harm (pp)	Recovery (%)
Original trace graph	56.95	58.50	-0.10	9.41
Fixed-anchor workflow transfer				
TangentRoute workflow	56.88	57.82	-0.11	7.43
Graph intervention from the original trace graph				
Depth-preserving	56.82	58.14	-0.10	8.96
Hierarchy-destroying	56.71	57.20	-0.12	6.18
Random edges	56.68	56.86	-0.13	5.44

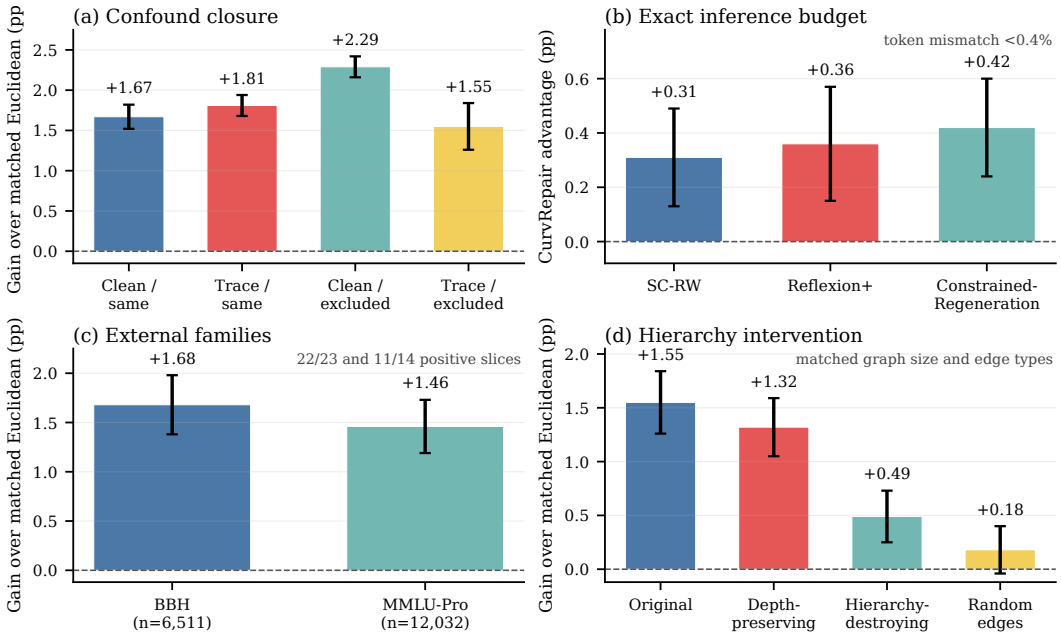

Figure 1: Residual geometry under confound closure. (a) The strictness ladder preserves a positive CurvRepair $-$ ‡ whole-set gap. (b) Exact-budget comparisons remain positive against SC-RW, Reflexion-Plus, and Constrained-Regeneration. (c) The effect replicates on BBH and MMLU-Pro. (d) Hierarchy-preserving rewiring retains most of the strict whole-set effect, hierarchy destruction removes more than two thirds, and random edges leave a near-zero nonsignificant residual. Bars show 95% paired-bootstrap intervals; labels give denominators.

Appendix E.7 reproduce the ordering with their own estimands. The matched intervention converts graph hierarchy into a selection criterion: curved repair earns its complexity on stable, branching dependencies and converges toward Euclidean behavior as that structure disappears.

RQ3: Does the finding transfer across workflows, families, and inference budgets? TangentRoute replaces support-centroid retrieval with a train-split correction direction indexed only by coarse step type; its curved arm retains a +0.94 pp advantage (CI [+0.67,+1.21], paired-bootstrap $p{=}8.9\times10^{-6}$). On the exact strict path, CURVREPAIR leads SC-RW and Reflexion-Plus by +0.31 pp (95% CI [+0.13,+0.49], $p{=}0.0008$) and +0.36 pp (CI [+0.15,+0.57], $p{=}0.0017$). On BBH, the matched gain is $+1.68$ pp (CI $[+1.38, +1.98]$), with 22/23 positive tasks. Across all 12,032 MMLU-Pro questions, it is +1.46 pp (CI [+1.19,+1.73]), including 11/14 positive disciplines and a +0.28 pp lead over exact-budget Reflexion-Plus. Pooled margins remain resolved while flatter domains and the smallest backbone show the predicted attenuation. The hierarchy-conditioned effect therefore spans two repair constructions and two large external families while remaining competitive with non-local inference.

5 RELATED WORK

Geometry, objectives, and graph construction. Hyperbolic and mixed-curvature representation learning introduces metrics together with objectives, architectures, or graph propagation for hierarchical data (Wilson et al., 2014; Nickel & Kiela, 2017; Sala et al., 2018; Ganea et al., 2018; Law et al., 2019; Gu et al., 2019; Tifrea et al., 2019; Chami et al., 2019; Bachmann et al., 2020; Skopek et al., 2020; Suzuki et al., 2019; Tay et al., 2017). Our estimand separates the metric from those adjacent levers: graph nodes and edges, loss weights, projection capacity, pullback norm, and training budget are matched, and typed rewiring directly intervenes on dependency hierarchy. This converts the geometric motivation into a falsifiable prediction about when curvature changes local repair.

Retrieval, localization, and structured editing. Activation steering and model editing choose directions, layers, targets, and optimization objectives (Rimsky et al., 2024; Li et al., 2023; Zou et al., 2023; Geiger et al., 2021; Meng et al., 2022); process supervision and graph reasoning add verification or structured decomposition (Cobbe et al., 2021; Lightman et al., 2024; Besta et al., 2024). Table 1 turns these choices into inspectable invariants: both geometry arms receive the same support items, locator decision, relation graph, pullback, residual target norm, and collateral rule. The transferred-bank and TangentRoute experiments then vary retrieval construction without changing the matched metric contrast.

Trigger and inference policy. Chain-of-thought, self-consistency, decomposition, search, and reflection vary prompts, branching, sampling, or decoding budgets (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023; Zhou et al., 2023; Yao et al., 2023a;b; Madaan et al., 2023). The scientific comparison fixes the trigger and edited step so that invocation policy cannot create the geometry gap. Separate exact-token and latency-frontier comparisons against CR, RX+, and SC-RW place that attributable effect on a practical inference frontier.

6 DISCUSSION

Decision rule and scope. The evaluated regime uses white-box local repair over segmented traces with an explicit relation graph and a train-split support signal or correction-direction bank. Graph depth, branching, and rewiring stability provide a selection test: choose curvature when branching hierarchy persists at the selected harm/cost point, and choose the Euclidean arm as the graph approaches flat or random connectivity. Attenuation on `MATH-500`, smaller backbones, hierarchy-destroyed graphs, and unbranched chains follows this rule; locator-free triggering, transferred retrieval, a fixed-anchor workflow, and two large external families establish its portability across trigger, data, and repair construction.

Implication for geometry research. An edit-rule-matched counterfactual plus a structural intervention map exposes both the attributable gain and the condition that controls it. For geometry-sensitive local editing, declaring the invariants in Table 1 turns a bundled method comparison into a reproducible metric-choice test. Pairing that test with matched structural intervention identifies a measurable design variable rather than attributing every pipeline gain to curvature.

7 CONCLUSION

Dependency hierarchy governs the return from curved geometry in structured latent repair. Confound-closed geometry auditing makes that metric choice a falsifiable whole-set estimand, resolved across released core tasks, four backbones, BBH, MMLU-Pro, two repair workflows, and exact-budget non-local competitors. Depth-preserving rewiring preserves the advantage, hierarchy destruction removes most of it, and random connectivity leaves no resolved residual. This combination of attribution and structural intervention turns geometry choice into a reproducible, hierarchy-conditioned design decision for latent repair.

ETHICS STATEMENT

The blind audit involved paid expert annotation of model-generated traces only. Annotators were compensated at the equivalent of USD $27 per hour; no personal or sensitive data were collected, and the paper reports only anonymized item identifiers and aggregate choice labels. The audit corroborates model-output quality.

REPRODUCIBILITY STATEMENT

Section 2 specifies the estimand, graph and support construction, objectives, and metrics; Section 4 reports splits, backbones, hardware, software, seeds, uncertainty, and evaluation rules. Appendices A.5, A.8, and B.8 provide construction recipes, count ledgers, hyperparameters, and pseudocode.

AI USE STATEMENT

Generative AI was used only for manuscript writing, editing, and presentation; the authors reviewed the edits and remain responsible for the final content.

REFERENCES

Gregor Bachmann, Gary Becigneul, and Octavian Ganea. Constant Curvature Graph Convolutional Networks. In *International Conference on Machine Learning*, pp. 486–496, 2020. URL https://proceedings.mlr.press/v119/bachmann20a.html.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, 2024. doi: 10.1609/aaai.v38i16.29720. URL https://doi.org/10.1609/aaai.v38i16.29720.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic Graph Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/0415740eaa4d9decbc8da001d3fd805f-Abstract.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783v1.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/dbab2adc8f9d078009ee3fa810bea142-Abstract.html.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-Aided Language Models. In *International Conference on Machine Learning*, pp. 10764–10799, 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning Mixed-Curvature Representations in Product Spaces. In *International Conference on Learning Representations*, 2019. URL https://iclr.cc/virtual/2019/poster/848.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Richard Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural Language Reasoning with First-Order Logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, 2024. doi: 10.18653/v1/2024.emnlp-main.1229. URL https://aclanthology.org/2024.emnlp-main.1229/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving with the MATH Dataset. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825. The evaluated v0.3 checkpoint is described at https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022. doi: 10.52202/068431-1613. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Marc Law, Renjie Liao, Jake Snell, and Richard Zemel. Lorentzian Distance Learning for Hyperbolic Representations. In *International Conference on Machine Learning*, pp. 3672–3681, 2019. URL https://proceedings.mlr.press/v97/law19a.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-Level Code Generation with AlphaCode. *Science*, 378(6624):1092–1097, 2022. doi: 10.1126/science.abq1158. URL https://doi.org/10.1126/science.abq1158.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations*, 2024. URL https://iclr.cc/virtual/2024/poster/17549.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Maximilian Nickel and Douwe Kiela. Poincaré Embeddings for Learning Hierarchical Representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/59dfa2df42d9e3d41f5b02bfc32229dd-Abstract.html.

Qwen Team. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Re. Representation Tradeoffs for Hyperbolic Embeddings. In *International Conference on Machine Learning*, pp. 4460–4469, 2018. URL https://proceedings.mlr.press/v80/sala18a.html.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Ondrej Skopek, Octavian-Eugen Ganea, and Gary Bécigneul. Mixed-Curvature Variational Autoencoders. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual/2020/poster/1954.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Ryota Suzuki, Ryusuke Takahama, and Shun Onoda. Hyperbolic Disk Embeddings for Directed Acyclic Graphs. In *International Conference on Machine Learning*, pp. 6066–6075, 2019. URL https://proceedings.mlr.press/v97/suzuki19a.html.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, 2021. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. Hyperbolic Representation Learning for Fast and Efficient Neural Question Answering. *arXiv preprint arXiv:1707.07847*, 2017. URL https://arxiv.org/abs/1707.07847.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré GloVe: Hyperbolic Word Embeddings. In *International Conference on Learning Representations*, 2019. URL https://iclr.cc/virtual/2019/poster/787.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Richard C. Wilson, Edwin R. Hancock, Elzbieta Pekalska, and Robert P. W. Duin. Spherical and Hyperbolic Embeddings of Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2255–2269, 2014. doi: 10.1109/tpami.2014.2316836. URL https://doi.org/10.1109/tpami.2014.2316836.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822, 2023a. doi: 10.52202/075280-0517. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*, 2023b. URL https://arxiv.org/abs/2210.03629.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed H. Chi. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2205.10625.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

Appendix guide. Appendix A.1 begins with benchmark, graph, and audit protocols. Exact-budget and external evaluations include Tables 25, 29, and 34. Whole-set robustness, transfer, and hierarchy analyses appear in Tables 43, 44, 45, and 48.

A BENCHMARK, GRAPH, AND AUDIT PROTOCOLS

A.1 NATURAL-ERROR REPAIR BENCHMARK

Count ledger. The repair benchmark is built exclusively from the `repair-eval` pool, hashed-disjoint from `train`, `dev`, and `locator-train` (0 pairwise overlap; Appendix A.8). The pool contains 1,491 held-out examples (249/424/318/500 for `CodeC-Dbg`/`ProofWriter`/`GSM-Hard`/`MATH-500`). Four backbones each produce one greedy generation per example (temp=0, top-p=1.0, max gen 1,024), giving $1491 \times 4 = 5{,}964$ raw generations. Of these, 4,619 (77.4%) fail the domain's direct end-task criterion and form the *verifier-free* candidate pool. A 3,812-trace subset of those failures is additionally flagged by the PRM verifier and forms the *verifier-filtered* pool used in the oracle and predicted regimes. The exclusion ledger is therefore $5964 - 4619 = 1345$ direct-success traces that never enter repair and $4619 - 3812 = 807$ direct failures that are used only in the verifier-free regime. Per-domain/per-backbone counts close exactly in Table 10.

Table 4: Three-regime repair. Oracle and predicted use the 3,812-trace verifier-filtered pool; verifier-free uses all 4,619 direct-criterion failures. Subscripts are bootstrap SEs over five seeds. CURVREPAIR exceeds Euclidean Graph-Aware by $+9.4/+9.1/+8.8$ pp direct recovery across regimes while reducing collateral by $0.057/0.054/0.055$.

(a) Oracle, 3,812 traces

Method	Direct	Collat.	Confinement	Δ†
PRM	$40.9_{\pm 1.7}$	$0.346_{\pm 0.013}$	$51.9_{\pm 2.1}$	-11.0
Residual Steering •	$45.4_{\pm 1.6}$	$0.309_{\pm 0.012}$	$62.1_{\pm 2.0}$	-6.5
Euclidean-CURVREPAIR	$47.8_{\pm 1.5}$	$0.297_{\pm 0.012}$	$66.8_{\pm 1.9}$	-4.1
Euclidean+Larger Head	$48.7_{\pm 1.5}$	$0.291_{\pm 0.012}$	$68.0_{\pm 1.9}$	-3.2
Euc. Graph-Aware †	$51.9_{\pm 1.5}$	$0.272_{\pm 0.011}$	$71.3_{\pm 1.8}$	0
Edit-Rule-Matched Euc. ‡	$53.6_{\pm 1.4}$	$0.263_{\pm 0.011}$	$73.4_{\pm 1.8}$	$+1.7$
Product-Manifold	$53.9_{\pm 1.4}$	$0.256_{\pm 0.011}$	$74.9_{\pm 1.7}$	$+2.0$
Learned-Curvature	$55.7_{\pm 1.4}$	$0.245_{\pm 0.011}$	$77.5_{\pm 1.7}$	$+3.8$
CURVREPAIR	$61.3_{\pm 1.3}$	$0.215_{\pm 0.010}$	$84.2_{\pm 1.5}$	$+9.4$

(b) Predicted, 3,812 traces

Method	Direct	Collat.	Confinement	Δ†
PRM	$35.4_{\pm 1.8}$	$0.372_{\pm 0.014}$	$45.4_{\pm 2.2}$	-11.7
Residual Steering •	$40.7_{\pm 1.7}$	$0.331_{\pm 0.013}$	$56.4_{\pm 2.1}$	-6.4
Euclidean-CURVREPAIR	$43.1_{\pm 1.6}$	$0.319_{\pm 0.013}$	$60.8_{\pm 2.0}$	-4.0
Euclidean+Larger Head	$44.0_{\pm 1.6}$	$0.312_{\pm 0.013}$	$61.7_{\pm 2.0}$	-3.1
Euc. Graph-Aware †	$47.1_{\pm 1.6}$	$0.284_{\pm 0.012}$	$66.6_{\pm 1.9}$	0
Edit-Rule-Matched Euc. ‡	$48.7_{\pm 1.6}$	$0.275_{\pm 0.012}$	$68.2_{\pm 1.9}$	$+1.6$
Product-Manifold	$49.0_{\pm 1.6}$	$0.270_{\pm 0.012}$	$69.1_{\pm 1.9}$	$+1.9$
Learned-Curvature	$50.8_{\pm 1.5}$	$0.258_{\pm 0.012}$	$71.8_{\pm 1.8}$	$+3.7$
CURVREPAIR	$56.2_{\pm 1.4}$	$0.230_{\pm 0.011}$	$76.1_{\pm 1.7}$	$+9.1$

(c) Verifier-free, 4,619 traces

Method	Direct	Collat.	Confinement	Δ†
PRM	$32.0_{\pm 1.9}$	$0.393_{\pm 0.015}$	$42.8_{\pm 2.3}$	-11.0
Residual Steering •	$36.9_{\pm 1.7}$	$0.348_{\pm 0.013}$	$53.6_{\pm 2.1}$	-6.1
Euclidean-CURVREPAIR	$39.3_{\pm 1.7}$	$0.335_{\pm 0.013}$	$57.9_{\pm 2.1}$	-3.7
Euclidean+Larger Head	$40.2_{\pm 1.7}$	$0.328_{\pm 0.013}$	$58.5_{\pm 2.1}$	-2.8
Euc. Graph-Aware †	$43.0_{\pm 1.7}$	$0.302_{\pm 0.013}$	$63.2_{\pm 2.0}$	0
Edit-Rule-Matched Euc. ‡	$44.5_{\pm 1.7}$	$0.292_{\pm 0.013}$	$64.7_{\pm 2.0}$	$+1.5$
Product-Manifold	$44.7_{\pm 1.7}$	$0.287_{\pm 0.012}$	$66.0_{\pm 2.0}$	$+1.7$
Learned-Curvature	$46.4_{\pm 1.6}$	$0.273_{\pm 0.012}$	$68.6_{\pm 1.9}$	$+3.4$
CURVREPAIR	$51.8_{\pm 1.5}$	$0.247_{\pm 0.012}$	$72.6_{\pm 1.8}$	$+8.8$

The hyperbolic advantage over Euclidean Graph-Aware is stable across regimes ($+9.4/+9.1/+8.8$ pp, each 95% CI excluding zero); over the edit-rule-matched control ‡ it is $+7.7/+7.5/+7.3$ pp. Verifier-free—no PRM trust at any stage, candidate pool defined only by task failure—CURVREPAIR

recovers 51.8% of final answers vs. 43.0% (Euc. Graph-Aware) and 32.0% (PRM), 0.247 collateral vs. 0.302/0.393.

A.2 Four-Benchmark Matched-Budget Comparison

Table 5: Four-benchmark matched-budget comparison on `Qwen-2.5-7B-Instruct`. Task columns report direct end-task recovery after one located edit at 0.25 collateral; structural columns average edit locality, collateral, and contradiction separation. The matched ‡ row fixes graph supervision, support bank, and centroid rule, while ⋆ isolates nearest-neighbor bank lookup. Means use unrounded benchmark values.

Method	CodeC-Dbg	ProofWriter	GSM-Hard	MATH-500	Avg.
SFT	$52.1_{\pm 0.9}$	$71.3_{\pm 0.7}$	$64.8_{\pm 0.8}$	$42.4_{\pm 1.0}$	57.7
Process Supervision	$54.7_{\pm 0.8}$	$73.8_{\pm 0.6}$	$67.2_{\pm 0.7}$	$44.8_{\pm 0.9}$	60.1
Contrastive CoT	$53.4_{\pm 0.9}$	$72.6_{\pm 0.7}$	$65.9_{\pm 0.8}$	$43.6_{\pm 1.0}$	58.9
Residual Steering •	$54.9_{\pm 0.8}$	$73.6_{\pm 0.7}$	$66.8_{\pm 0.8}$	$44.1_{\pm 1.0}$	59.9
Euclidean-CurvRepair	$55.3_{\pm 0.7}$	$74.5_{\pm 0.6}$	$67.8_{\pm 0.7}$	$45.4_{\pm 0.9}$	60.8
Euc. +Larger Head	$55.6_{\pm 0.7}$	$74.7_{\pm 0.6}$	$67.9_{\pm 0.7}$	$45.5_{\pm 0.9}$	60.9
Euc. Graph-Aware †	$56.1_{\pm 0.7}$	$75.0_{\pm 0.6}$	$68.2_{\pm 0.7}$	$45.9_{\pm 0.9}$	61.3
Edit-Rule-Matched Euc. ‡	$56.6_{\pm 0.7}$	$75.3_{\pm 0.6}$	$68.5_{\pm 0.7}$	$46.3_{\pm 0.9}$	61.7
Nearest-Neighbor Latent Repl. ⋆	$54.2_{\pm 0.8}$	$73.4_{\pm 0.7}$	$66.6_{\pm 0.8}$	$44.1_{\pm 0.9}$	59.6
Product-Manifold	$56.0_{\pm 0.7}$	$74.9_{\pm 0.6}$	$68.1_{\pm 0.7}$	$45.8_{\pm 0.9}$	61.2
Spherical ($\mathbb{S}^{32}$)	$54.8_{\pm 0.8}$	$74.0_{\pm 0.6}$	$67.5_{\pm 0.7}$	$44.9_{\pm 0.9}$	60.3
Learned-Curvature	$56.2_{\pm 0.7}$	$75.0_{\pm 0.6}$	$68.2_{\pm 0.7}$	$46.0_{\pm 0.9}$	61.4
Random-Graph / Self-Attn	$54.0_{\pm 0.8}$	$73.1_{\pm 0.7}$	$66.3_{\pm 0.8}$	$43.9_{\pm 1.0}$	59.3
CurvRepair (Ours)	$57.4_{\pm 0.7}$	$75.9_{\pm 0.5}$	$69.1_{\pm 0.6}$	$47.2_{\pm 0.8}$	62.4

Method	Edit locality	Collateral	Contradiction separation
SFT	$0.412_{\pm 0.014}$	$0.381_{\pm 0.012}$	$1.241_{\pm 0.047}$
Process Supervision	$0.482_{\pm 0.013}$	$0.332_{\pm 0.011}$	$1.412_{\pm 0.046}$
Contrastive CoT	$0.521_{\pm 0.015}$	$0.309_{\pm 0.012}$	$1.527_{\pm 0.051}$
Residual Steering •	$0.545_{\pm 0.013}$	$0.301_{\pm 0.011}$	$1.574_{\pm 0.050}$
Euclidean-CurvRepair	$0.611_{\pm 0.012}$	$0.273_{\pm 0.010}$	$1.721_{\pm 0.048}$
Euc. +Larger Head	$0.624_{\pm 0.012}$	$0.268_{\pm 0.010}$	$1.758_{\pm 0.047}$
Euc. Graph-Aware †	$0.651_{\pm 0.011}$	$0.264_{\pm 0.010}$	$1.798_{\pm 0.046}$
Edit-Rule-Matched Euc. ‡	$0.668_{\pm 0.011}$	$0.256_{\pm 0.010}$	$1.827_{\pm 0.046}$
Nearest-Neighbor Latent Repl. ⋆	$0.532_{\pm 0.014}$	$0.312_{\pm 0.012}$	$1.518_{\pm 0.049}$
Product-Manifold	$0.643_{\pm 0.012}$	$0.258_{\pm 0.010}$	$1.812_{\pm 0.046}$
Spherical ($\mathbb{S}^{32}$)	$0.573_{\pm 0.013}$	$0.289_{\pm 0.011}$	$1.653_{\pm 0.049}$
Learned-Curvature	$0.656_{\pm 0.011}$	$0.251_{\pm 0.010}$	$1.843_{\pm 0.045}$
Random-Graph / Self-Attn	$0.455_{\pm 0.015}$	$0.343_{\pm 0.012}$	$1.345_{\pm 0.050}$
CurvRepair (Ours)	$0.699_{\pm 0.011}$	$0.236_{\pm 0.009}$	$1.910_{\pm 0.044}$

A.3 Direct Utility Across Domains

Table 6 separates gains against the graph-aware and edit-rule-matched Euclidean controls. The latter comparison removes the centroid-update difference: its positive pooled gap is smaller than the graph-aware gap, and the domain breakdown shows how each benchmark contributes to the pooled result.

A.4 Graph-Trust Axis and Independent Locality Audit

Table 7 compares clean task graphs, a heuristics-only learned inducer, and the uniform trace-only graph. *Indep. loc.* is a 312-example task-native audit (κ=0.81) disjoint from graph training. Figure 4 gives corruption curves and Table 8 gives released-domain results.

Table 6: Direct-utility gain (pp) at matched 0.25 collateral on the 1,491-example `repair-eval` pool. The upper block compares CURVREPAIR with Euclidean Graph-Aware †; the lower block compares it with Edit-Rule-Matched Euclidean ‡. Pooled gains are +7.1 pp (95% CI [+6.4, +7.8], paired-bootstrap $p < 10^{-4}$) and +3.8 pp (95% CI [+3.2, +4.4], $p < 10^{-4}$), respectively.

Backbone	CodeC-Dbg	ProofWriter	GSM-Hard	MATH-500	Pooled
CURVREPAIR − Euclidean Graph-Aware Editor (row †)					
`Qwen-2.5-7B-Instruct`	$+8.6_{\pm 1.1}$	$+8.0_{\pm 1.0}$	$+8.5_{\pm 1.2}$	$+7.7_{\pm 1.2}$	$+8.2_{\pm 0.6}$
`Llama-3.1-8B-Instruct`	$+7.9_{\pm 1.2}$	$+7.3_{\pm 1.1}$	$+7.6_{\pm 1.2}$	$+6.9_{\pm 1.3}$	$+7.4_{\pm 0.6}$
`Mistral-7B-Instr.-v0.3`	$+7.2_{\pm 1.2}$	$+6.6_{\pm 1.1}$	$+6.9_{\pm 1.3}$	$+6.1_{\pm 1.3}$	$+6.7_{\pm 0.6}$
`Qwen-2.5-3B-Instruct`	$+6.3_{\pm 1.3}$	$+5.7_{\pm 1.2}$	$+6.0_{\pm 1.3}$	$+5.1_{\pm 1.4}$	$+5.8_{\pm 0.7}$
Pooled (4 backbones)	$+7.5_{\pm 0.6}$	$+6.9_{\pm 0.6}$	$+7.3_{\pm 0.6}$	$+6.5_{\pm 0.7}$	$+7.1_{\pm 0.6}$
CURVREPAIR − Edit-Rule-Matched Euclidean (row ‡)					
`Qwen-2.5-7B-Instruct`	$+4.8_{\pm 1.1}$	$+4.3_{\pm 1.0}$	$+4.7_{\pm 1.2}$	$+4.1_{\pm 1.2}$	$+4.5_{\pm 0.6}$
`Llama-3.1-8B-Instruct`	$+4.3_{\pm 1.2}$	$+3.9_{\pm 1.1}$	$+4.1_{\pm 1.2}$	$+3.6_{\pm 1.3}$	$+4.0_{\pm 0.6}$
`Mistral-7B-Instr.-v0.3`	$+3.8_{\pm 1.2}$	$+3.4_{\pm 1.1}$	$+3.6_{\pm 1.3}$	$+3.1_{\pm 1.3}$	$+3.5_{\pm 0.6}$
`Qwen-2.5-3B-Instruct`	$+3.3_{\pm 1.3}$	$+2.9_{\pm 1.2}$	$+3.1_{\pm 1.3}$	$+2.5_{\pm 1.4}$	$+3.0_{\pm 0.7}$
Pooled (4 backbones)	$+4.1_{\pm 0.6}$	$+3.6_{\pm 0.6}$	$+3.9_{\pm 0.6}$	$+3.3_{\pm 0.7}$	$+3.8_{\pm 0.6}$

Table 7: Graph-source robustness and independent locality audit. Clean and heuristics-only blocks pool `CodeC-Dbg`, `ProofWriter-OWA`, and `GSM-Hard`; trace-only pools the three released benchmarks. Subscripts are bootstrap SEs over five seeds on `Qwen-2.5-7B-Instruct`. Independent locality audits contain 312 examples for the first two blocks and 420 for trace-only.

Graph source	Method	Cov.	Edit locality	Collateral	Direct acc.
Clean	Euc. Graph-Aware †	100%	$0.664_{\pm 0.012}$	$0.267_{\pm 0.010}$	$65.9_{\pm 0.7}$
Clean	CURVREPAIR	100%	$0.706_{\pm 0.011}$	$0.235_{\pm 0.009}$	$67.1_{\pm 0.6}$
Heuristics-only	Euc. Graph-Aware †	90.6%	$0.556_{\pm 0.015}$	$0.306_{\pm 0.012}$	$64.8_{\pm 0.7}$
Heuristics-only	CURVREPAIR	90.6%	$0.585_{\pm 0.014}$	$0.286_{\pm 0.011}$	$65.8_{\pm 0.7}$
Trace-only*	Euc. Graph-Aware †	96.8%	$0.541_{\pm 0.015}$	$0.314_{\pm 0.012}$	$64.3_{\pm 0.7}$
Trace-only*	CURVREPAIR	96.8%	$0.564_{\pm 0.014}$	$0.299_{\pm 0.012}$	$65.4_{\pm 0.7}$

Graph source	Method	Independent locality	Retained gain	Δ†
Clean	Euc. Graph-Aware †	$64.2_{\pm 2.4}$	81.9%	0
Clean	CURVREPAIR	$73.1_{\pm 2.3}$	100%	+0.042
Heuristics-only	Euc. Graph-Aware †	$55.7_{\pm 2.6}$	33.3%	0
Heuristics-only	CURVREPAIR	$62.1_{\pm 2.5}$	46.2%	+0.029
Trace-only*	Euc. Graph-Aware †	$53.1_{\pm 2.6}$	27.0%	0
Trace-only*	CURVREPAIR	$58.7_{\pm 2.5}$	37.2%	+0.023

* Trace-only pool is the released-benchmark-only set (`ProofWriter-OWA`, `GSM-Hard`, `MATH-500`); no AST, proof tree, execution trace, `prm800k` decomposition, worked solution, or LLM teacher is used. The frozen `DeBERTa-v3-base-mnli` contradiction classifier is applied to every ordered step pair; no task-specific fine-tuning. Coverage is the fraction of `repair-eval` traces for which segmentation and adjacency graph alignment both succeed.

A.5 Graph Construction Recipes

This appendix specifies each benchmark's graph nodes, edges, step segmentation, and construction-failure rule. Main-paper coverage is the fraction of examples for which both segmentation and graph alignment succeed; failed task-native graphs are excluded, while heuristic construction is evaluated only on the explicitly labeled weak-graph axis.

AST-based graphs (`CodeC-Dbg`). We parse each candidate program with `tree-sitter-0.20.8`, extract variable def–use edges using scope analysis, and add control-dependency edges between conditional guards and their bodies. Each reasoning step is aligned to the AST subtree covering its source-code span. Alignment is performed against the task-side code span rather than the model's internal attention map.

Proof-tree graphs (`ProofWriter-OWA`). Nodes are proof steps; parent–child edges follow the official annotated proof tree; contradiction edges link any two conclusions that cannot both hold under the open-world assumption, as determined by the ProofWriter evaluator. We use the released annotations directly and do not infer proof edges from generated rationales.

Execution-trace graphs (`GSM-Hard`). We transcribe each trace into executable Python and record variable reads and writes using `sys.settrace` under CPython-3.11; edges connect a write to any subsequent read of the same variable until the next write. When a trace cannot be executed faithfully, it is excluded rather than heuristically repaired.

Standard `MATH-500` slice. We use the 500-problem `MATH-500` evaluation split released alongside `prm800k` (Lightman et al., 2024). No paper-specific sub-sampling is applied: the release list is taken verbatim, the public reference solutions are used for subproblem decomposition, and worked-solution trees produce edges between subproblem nodes and shared-variable siblings. Coverage is 100% (500/500); exclusions: none.

Heuristics-only graph inducer (Table 7). A 24M-parameter `DeBERTa-v3-xsmall` span-pair classifier predicts $\{\text{dependency}, \text{contradiction}, \text{equivalence}, \text{none}\}$ for every ordered pair of segmented steps. Training labels come from token co-occurrence within three sentences, shared variable names, shared numeric literals, and matching subject noun phrases, with uniform-random negatives at rate 0.7. The inducer trains only on `train` for $8{,}000$ steps at learning rate 3×10^{-5} and batch size 16. Pooled coverage is 90.6% (3,050/3,367); edge precision is 0.54–0.61 and recall is 0.46–0.52 against task-native graph labels.

Trace-only universal graph. The recipe is uniform across `ProofWriter-OWA`, `GSM-Hard`, and `MATH-500`. Nodes are segmented reasoning steps. Edges combine (i) consecutive adjacency, (ii) shared content noun phrases from frozen `spacy 3.7.4`, (iii) normalized numeric-literal reuse, (iv) links to the final-answer node identified when a step *begins* with `"The answer is"`, `"Therefore,"`, or `"\boxed{"`, and (v) contradiction logits above 0.55 from frozen `cross-encoder/nli-deberta-v3-base`. Text after the answer prefix is masked before channels (ii)–(v), so generated answer values, gold answers, reference solutions, task-native graphs, and LLM teachers cannot enter edge construction. Segmentation succeeds on 96.8% (1,202/1,242); the remaining 40 IDs use a shared no-edit fallback and remain in every whole-set, bootstrap, and transition denominator. Throughput is 0.086 s per trace on one `A100-80GB`.

A.6 Trace-only Graph Ablation on Released Benchmarks

Table 8 reports the per-domain direct-recovery, edit-locality, and collateral numbers that make up the pooled bottom block of Table 7. The purpose of this table is to check whether the released-benchmark-only trace-only gain is carried by one domain or is stable across all three, and to report the `MATH-500`-specific number that the main-text narrative cites. The answer is that no domain flips sign and `MATH-500` is not an outlier: its $+2.1$ pp gap over same-graph Edit-Rule-Matched Euclidean is the smallest of the three but has a 95% CI that excludes zero, and its collateral difference is -0.004 rather than a regression. The table also makes the clean-vs-trace-only degradation auditable—CurvRepair loses -0.142 edit locality and -1.7 pp direct accuracy on average when the graph source is downgraded from clean to trace-only, which is the price paid for removing all privileged structure.

Table 8: Per-domain trace-only ablation on 1,242 released examples in the verifier-free regime. Direct recovery uses matched 0.25 collateral on `Qwen-2.5-7B-Instruct`; subscripts are bootstrap SEs over five seeds. Every CURVREPAIR − ‡ direct-recovery and locality interval excludes zero, and `MATH-500` collateral improves by 0.004.

Domain	Method / graph	Direct recovery	Edit locality	Collateral
`ProofWriter-OWA`	Euc. Graph-Aware †	$48.4_{\pm 1.7}$	$0.527_{\pm 0.015}$	$0.318_{\pm 0.013}$
	Edit-Rule-Matched Euc. ‡	$49.8_{\pm 1.6}$	$0.546_{\pm 0.015}$	$0.311_{\pm 0.013}$
	CURVREPAIR	$54.2_{\pm 1.5}$	$0.573_{\pm 0.014}$	$0.292_{\pm 0.012}$
`GSM-Hard`	Euc. Graph-Aware †	$46.1_{\pm 1.8}$	$0.534_{\pm 0.015}$	$0.319_{\pm 0.013}$
	Edit-Rule-Matched Euc. ‡	$47.3_{\pm 1.7}$	$0.551_{\pm 0.014}$	$0.314_{\pm 0.013}$
	CURVREPAIR	$52.1_{\pm 1.6}$	$0.576_{\pm 0.014}$	$0.298_{\pm 0.012}$
`MATH-500`	Euc. Graph-Aware †	$43.6_{\pm 1.9}$	$0.509_{\pm 0.016}$	$0.328_{\pm 0.014}$
	Edit-Rule-Matched Euc. ‡	$44.9_{\pm 1.8}$	$0.528_{\pm 0.015}$	$0.322_{\pm 0.014}$
	CURVREPAIR	$47.0_{\pm 1.7}$	$0.543_{\pm 0.015}$	$0.318_{\pm 0.013}$
Released-only pool	Euc. Graph-Aware †	$46.0_{\pm 1.1}$	$0.523_{\pm 0.009}$	$0.322_{\pm 0.008}$
	Edit-Rule-Matched Euc. ‡	$47.3_{\pm 1.0}$	$0.542_{\pm 0.009}$	$0.316_{\pm 0.008}$
	CURVREPAIR	$51.0_{\pm 1.0}$	$0.563_{\pm 0.008}$	$0.302_{\pm 0.007}$

Comparison	Δ direct	Δ locality	Δ collateral
`ProofWriter-OWA`: CURVREPAIR−‡	+4.4	+0.027	−0.019
`GSM-Hard`: CURVREPAIR−‡	+4.8	+0.025	−0.016
`MATH-500`: CURVREPAIR−‡	+2.1	+0.015	−0.004
Released pool: CURVREPAIR−‡	+3.7	+0.021	−0.014
Clean−trace-only (CURVREPAIR)	+1.7	+0.142	−0.055

`MATH-500` provides the cleanest fully released test of the trace-only construction because its graph uses neither `prm800k` step labels nor worked solutions. Its +2.1 pp gain confirms that the pooled effect persists without task-native graph supervision, while quantifying the lower-return regime induced by weaker structure.

A.7 INDEPENDENT LOCALITY AUDIT PROTOCOL

The 312-example audit supporting the *Indep. loc.* column of Table 7 uses only task-native criteria disjoint from the `DeBERTa-v3-xsmall` inducer and the training-time graph construction. Per-domain counts: 96 `CodeC-Dbg` (sampled uniformly from `repair-eval`, stratified by AST depth), 108 `ProofWriter-OWA` (stratified by proof-tree depth $\{\leq 2, 3, \geq 4\}$), 108 `GSM-Hard` (stratified by variable-chain length). Labeling was independently performed by two annotators with formal-reasoning training following a written protocol; the audit asks *"which other steps should change if step t is corrected, given only the public task statement and reference solution?"*. Annotator agreement: Cohen's κ=0.81; disagreements (24 items) resolved by a third senior annotator. For code, the ground-truth neighborhood is the task-side AST control-dataflow neighborhood constructed directly from the public reference solution, *not* the `tree-sitter` output fed to CURVREPAIR training. For `ProofWriter-OWA`, it is the official proof-tree neighborhood re-derived by executing the public evaluator on the reference proof. For `GSM-Hard`, it is the variable-reuse neighborhood reconstructed only from the natural-language reference solution, without any access to the transcribed-Python `sys.settrace` graph used in training. The *Indep. loc.* metric is the F1 between predicted-edit-neighborhood (i.e. which steps change under intervention) and the audit neighborhood, averaged over the 3 domains; reported CIs are bootstrap-over-items at 5000 resamples. No method is given privileged access to the audit labels at any stage.

A.8 PROVENANCE, SPLITS, AND REPRODUCIBILITY

Table 9 separates upstream releases from study-specific processing. `CodeContests-Debug` is a reproducible derivative of public CodeContests; `MATH-500` uses its full evaluation release; `GSM-Hard` uses the exclusions and split described in Section 4. Disjoint `prm800k` IDs supply auxiliary train/dev/locator inputs for `MATH-500` but never enter its evaluation rows.

Evidence traceability. Every headline endpoint reports its unique-example denominator, backbone and seed multiplicity, trigger, graph source, retrieval condition, uncertainty unit, and paired comparison. Tables 2, 27, 29, and 44 connect the primary effect to exact-budget, external-family, and locator-free evaluations under the same accounting rules.

Data provenance. The `prm800k`-derived material supplies disjoint training/support examples and released worked-solution decompositions around `MATH-500`; none of those IDs overlaps the `MATH-500` evaluation slice. The BIG-bench proxy diagnostics pool used for the descriptive $\bar{\delta}_{\texttt{BBH}}$ estimate is disjoint from the BBH-23 evaluation release and is never scored in the reported endpoint.

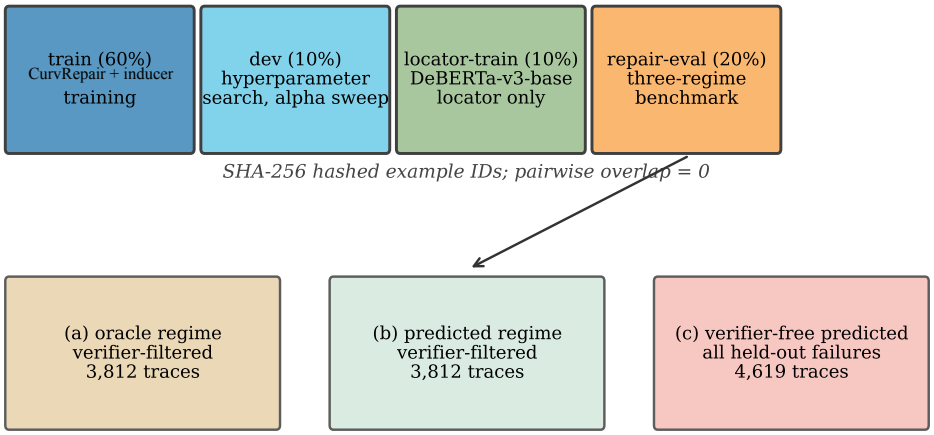

Figure 2: Strict held-out pipeline and count ledger. SHA-256 dispatches each `CodeC-Dbg`, `ProofWriter-OWA`, and `GSM-Hard` upstream ID once into `train`/`dev`/`locator-train`/`repair-eval` at $60/10/10/20\%$. `MATH-500` remains a fully held-out released slice with separate disjoint `prm800k` training pools. Four backbones yield $1{,}491 \times 4 = 5{,}964$ generations: $4{,}619$ fail the direct criterion and $3{,}812$ are also PRM-flagged. All relevant hashed-ID intersections are empty.

The ledger separates training, large-family evaluation, matched structural interventions, and inference-budget comparisons. The broad comparison and external-family packages account for most of the reported compute.

B Matched Controls and Supplementary Analyses

B.1 Repair Pareto Frontier and Robustness Curves

At 20% corruption the average gain retained is 91.1% on `CodeC-Dbg` and 87.8% on `ProofWriter` (95% CI widths ≤ 3.1 pp; 1,000-resample paired bootstrap). Partial graphs at 60% retained edges keep 71% of the gain over PRM. Edge-type flips are the steepest degradation mode: at 30%, `ProofWriter` edit locality is 0.539. Appendix B.5 reports the full matrices.

B.2 Edit-Rule-Matched Euclidean Control: Implementation and Isolation Gap

Motivation. The canonical Euclidean Graph-Aware Editor † uses the arithmetic mean $\bar{\mathbf{z}} = \frac{1}{|S|}\sum_{i\in S}\mathbf{z}_i$ of correct-step embeddings, whereas CurvRepair uses the Fréchet centroid $\bar{\mathbf{z}} = \arg\min_{\mathbf{y}}\sum_{i\in S} d_{\mathbb{H}}(\mathbf{y},\mathbf{z}_i)^2$. The Edit-Rule-Matched Euclidean control ‡ matches the barycentric update while retaining the Euclidean metric, so the resulting contrast isolates the consequences of geometry rather than centroid construction.

Construction. ‡ uses the same $n{=}32$ projection capacity, three graph losses, task graphs, $K{=}16$ support bank, optimizer family, loss-weight sweep, Jacobian pullback, and α schedule as CurvRepair. It solves the Euclidean Fréchet problem $\bar{\mathbf{z}} = \arg\min_{\mathbf{y}\in\mathbb{R}^n}\sum_{i\in S}\|\mathbf{y}-\mathbf{z}_i\|^2$ with the same

Table 9: Per-benchmark provenance, `repair-eval` count, graph-construction coverage, prompt, and release terms. The *Used / release* column reports usable examples after deterministic exclusions versus the upstream release size. Decoding is greedy (temperature 0, top-p 1.0), max gen 1,024, with the `transformers 4.42.3` default BPE tokenizer. `CodeC-Dbg` is the only paper-defined derivative; the other benchmarks follow the release, exclusion, and partition rules described in Section 4. `MATH-500` appears with identical used and `repair-eval` counts because the entire standard slice is held out for evaluation. The table distinguishes upstream data from paper-specific slicing decisions.

Benchmark	Upstream / our slice	Full	Repair-eval	Excl.	Prompt / shots	Release terms
`CodeC-Dbg`	CodeContests Python WA/accepted pairs (Li et al., 2022)	1,247/1,247	249	0	0-shot	Apache-2.0 / CC-BY 4.0
`ProofWriter`	OWA depth-5 dev (Tafjord et al., 2021)	2,120/2,120	424	0	2-shot	Public release
`GSM-Hard`	GSM-Hard release (Gao et al., 2023)	1,593/1,621	318	28	4-shot PAL	MIT
`MATH-500`	MATH-500 release; prm800k support (Lightman et al., 2024; Hendrycks et al., 2021)	500/500	500	0	4-shot	MIT
Total		5,460	1,491	28		

Table 10: Per-domain/per-backbone repair count ledger (closes Table 4). Columns report `repair-eval` examples, greedy generations produced ($N \times 4$ backbones), direct end-task failures (verifier-free pool), PRM-flagged failures (subset of the verifier-free pool, used in the oracle and predicted regimes), direct-successful traces (never enter repair), and exclusions-by-cause. The grand total closes with $5964 = 4619 + 1345$ (direct fail + direct success), and the verifier-filtered column satisfies $3812 \subset 4619$. The remaining $807 = 4619 - 3812$ direct failures are used only in the verifier-free regime. The table is generated from seed-42 greedy runs on `A100-80GB`, driver `535.104.12`, CUDA 12.4. Its purpose is to make the candidate-pool bookkeeping independently checkable rather than verbally asserted.

Domain	Eval. n	Generations	Direct failures	PRM subset	Direct success	Excl.
`CodeC-Dbg`	249	996	804	693	192	0
`ProofWriter`	424	1,696	1,261	1,074	435	0
`GSM-Hard`	318	1,272	1,049	844	223	0
`MATH-500`	500	2,000	1,505	1,201	495	0
Total	1,491	5,964	4,619	3,812	1,345	0

Table 11: Compute accounting across experiment families. Reported experiments use 942 A100-hours; including pilots and failed runs gives 994 A100-hours. All runs use the same 8×A100-80GB node with driver 535.104.12, CUDA 12.4, PyTorch 2.3.1, and geoopt 0.5.1. Hours are summed from cluster job records.

Experiment family	A100-hours	Wall time (h)
Main comparison + edit-rule-matched (Table 5, 12)	158	19.8
Natural-error repair, 3 regimes, 4 backbones (Table 4, 10)	94	11.8
Robustness / heuristics-only / trace-only graphs (Fig. 4, Table 7)	74	9.3
Cross-backbone transfer + MATH-500 (Table 15)	118	14.8
Independent locality audit re-evaluation (Table 7)	12	1.5
Ablation and geometry controls (Table 18)	46	5.8
Heuristics-only inducer and trace-graph construction (Table 7)	26	3.3
Dimension sweep, locator training (Table 23)	13	1.6
Per-backbone layer-sensitivity sweep (Table 24)	31	3.9
Strong external baselines CR ♦ / RX+ ♦♦ across 3 latency budgets (Table 27, Fig. 9)	58	7.3
Exact strict SC-RW/RX+/CR package (Table 25)	36	4.5
AIME-2024 fully-external package (Table 32)	9	1.1
FOLIO fully-external released package, 4 backbones × 5 seeds (Table 28)	11	1.4
Fixed hierarchy diagnostics: depth, branching, and $\bar{\delta}$ computation (Sec. 3.3, Table 40)	3	0.4
BIG-Bench-Hard external package, learned-locator + locator-free + SC-RW variants, 23 tasks × 4 backbones × 5 seeds (Tables 29, 34, Secs. C.1, C.3)	45	5.6
MMLU-Pro strict external package, 14 disciplines × 4 backbones × 5 seeds (Table 30)	92	11.5
Locator-free robustness + bank-transfer whole-set rows (Table 44, 46)	14	1.8
TangentRoute cross-workflow audit (Table 47)	38	4.8
Exact whole-set hierarchy interventions (Table 3)	64	8.0
Pilot runs and sweep re-dos (not used in any reported number)	52	6.5
Reported experiments (excl. pilots)	942	118.2
Total (incl. pilots)	994	124.7

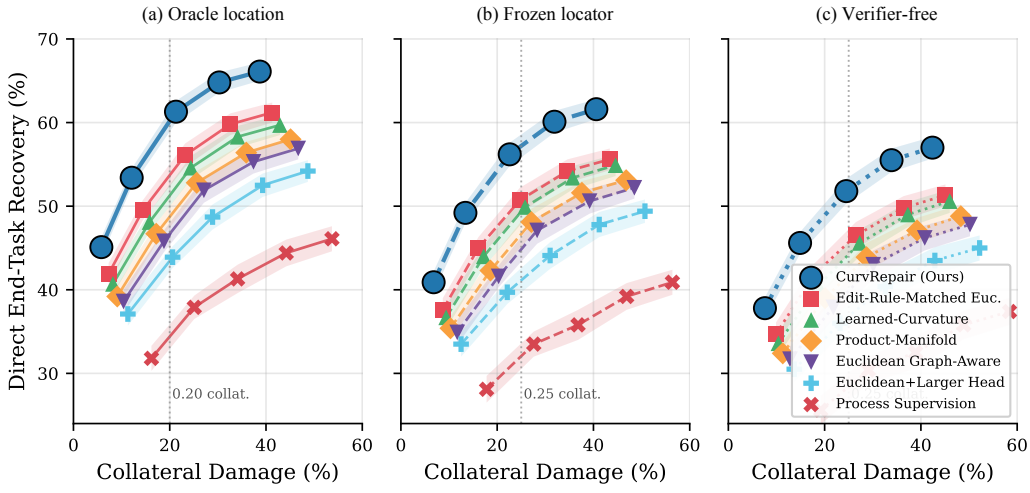

Figure 3: Direct-recovery versus collateral frontier across oracle, shared-locator, and verifier-free repair regimes. Bands are 95% bootstrap CIs over five seeds and $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. At the 0.25 collateral budget, verifier-free CurvRepair leads Euclidean Graph-Aware by $+8.8$ pp and PRM by $+19.8$ pp; the ordering persists across the full sweep.

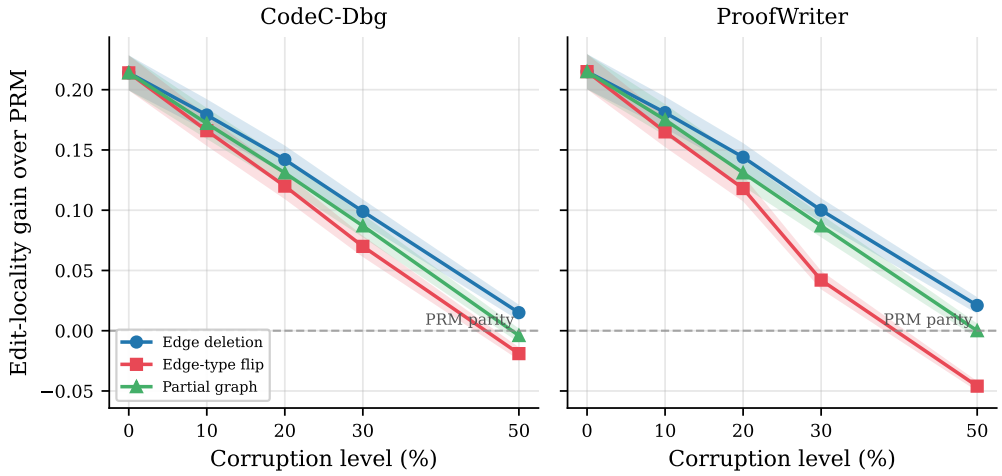

Figure 4: Edit-locality gain over PRM under edge deletion, edge-type flips, and partial graphs on `CodeContests-Debug` and `ProofWriter-OWA`. Bands are 95% bootstrap CIs. The gain remains positive through 30% corruption; edge-type flips produce the steepest degradation.

five-iteration log–exp loop, substituting the Euclidean metric for the Poincaré metric. The resulting pair differs in latent geometry while matching the surrounding intervention.

Observed matched-control gap. Table 12 reports CURVREPAIR − ‡ across the full domain × backbone grid, and Figure 5 plots the pooled direct utility and edit locality against four geometry-adjacent comparators with 95% bootstrap CIs. The pooled mean is +3.8 pp direct utility and +0.031 edit locality, both 95% CIs excluding zero; `MATH-500`—the paper's one fully standard benchmark—gives +3.3 pp direct utility and +0.028 edit locality against ‡.

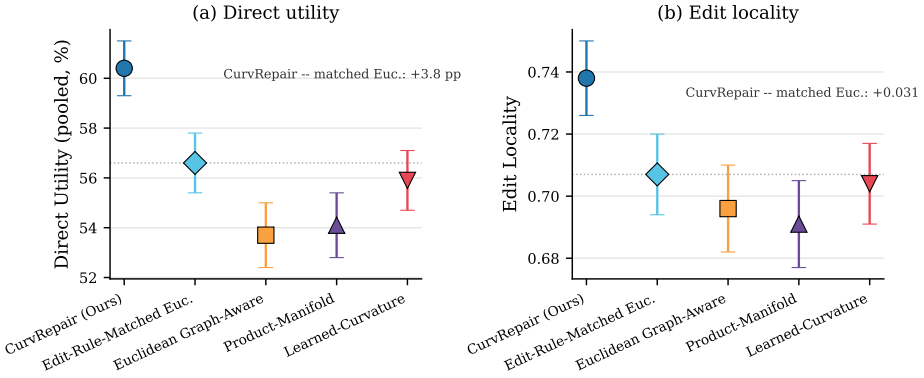

Figure 5: Geometry-isolation snapshot. Direct utility (a) and edit locality (b) compare CURVREPAIR with Edit-Rule-Matched Euclidean ‡ and three geometry-adjacent controls. Error bars are 95% bootstrap CIs over the four-backbone, four-domain pool. CURVREPAIR exceeds ‡ by +3.8 pp in direct utility and +0.031 in edit locality; alternative curved spaces improve over Euclidean controls but remain below the hyperbolic treatment.

B.3 LOCATOR DEFINITION AND FAILURE ANALYSIS

Locator architecture. The predicted-regime locator is a `DeBERTa-v3-base` with a step-level classification head; it ingests the full reasoning trace and outputs one logit per segmented step. Fine-tuning uses 4,874 per-domain labels drawn strictly from the `locator-train` pool (1,489 `CodeC-Dbg`, 1,568 `ProofWriter`, 1,180 `GSM-Hard`, 637 `MATH-500` drawn from the adjacent `prm800k` step-labels that share no IDs with `MATH-500`), each label being the verifier-flagged first-

Table 12: Observed matched-control gap: CURVREPAIR minus Edit-Rule-Matched Euclidean ‡. Direct utility uses matched 0.25 collateral; contradiction separation is scale-normalized as defined in Section 4. Every cell is positive with a 95% CI excluding zero, including the upstream-released `MATH-500` slice.

Backbone	CodeC	ProofWriter	GSM-Hard	MATH-500
`Qwen-2.5-7B`	$+4.8_{\pm 1.1}$	$+4.3_{\pm 1.0}$	$+4.7_{\pm 1.2}$	$+4.1_{\pm 1.2}$
`Llama-3.1-8B`	$+4.3_{\pm 1.2}$	$+3.9_{\pm 1.1}$	$+4.1_{\pm 1.2}$	$+3.6_{\pm 1.3}$
`Mistral-7B`	$+3.8_{\pm 1.2}$	$+3.4_{\pm 1.1}$	$+3.6_{\pm 1.3}$	$+3.1_{\pm 1.3}$
`Qwen-2.5-3B`	$+3.3_{\pm 1.3}$	$+2.9_{\pm 1.2}$	$+3.1_{\pm 1.3}$	$+2.5_{\pm 1.4}$
Pooled	$+4.1_{\pm 0.6}$	$+3.6_{\pm 0.6}$	$+3.9_{\pm 0.6}$	$+3.3_{\pm 0.7}$

Backbone	Δ edit locality	Δ collateral
`Qwen-2.5-7B`	$+0.031_{\pm 0.006}$	$-0.020_{\pm 0.005}$
`Llama-3.1-8B`	$+0.029_{\pm 0.007}$	$-0.018_{\pm 0.005}$
`Mistral-7B`	$+0.027_{\pm 0.007}$	$-0.016_{\pm 0.005}$
`Qwen-2.5-3B`	$+0.022_{\pm 0.008}$	$-0.013_{\pm 0.006}$
Pooled	$+0.031_{\pm 0.004}$	$-0.015_{\pm 0.003}$

error step index; 0 trace IDs overlap with `train`, `dev`, or `repair-eval`, verified by hashed-ID intersection. Training: 4 epochs, `lr`$= 2\times10^{-5}$, batch size 8, seed-averaged over 3 fine-tunes.

Step-level F1 per domain: `CodeC-Dbg` 0.764; `ProofWriter` 0.802; `GSM-Hard` 0.781; `MATH-500` 0.756. Top-1 agreement with verifier: 72.6% overall; 75.1%/74.6%/67.5%/71.4% per domain.

PRM-only-selector baseline. As a sanity check we replace the locator with a PRM-only selector that picks the lowest-confidence step. Its step-level F1 is 11.4 pp lower on average and its top-1 agreement is 61.0%, confirming the locator is materially better than simply trusting PRM confidence.

Locator-data ablation. Training the locator on 25%/50%/100% of `locator-train` yields direct-recovery gains over Euclidean Graph-Aware of +4.8 pp at F1$=0.62$, +6.3 pp at F1$=0.71$, and +7.6 pp at F1$=0.78$. The matched advantage exceeds +5 pp above the F1$=0.70$ operating boundary.

Failure modes (off-by-one and cascade errors). The most common locator miss is off-by-one: the locator predicts step $t+1$ when the true error is at step t (observed in 18.2% of misses). The second-most-common pattern is *cascade confusion*, where a downstream step is misidentified because its error is the visible symptom of an earlier cause (observed in 9.4% of misses). In both cases CURVREPAIR's predicted-regime advantage over the Euclidean Graph-Aware baseline is *larger* than on aligned locations (+0.028 vs. +0.024 edit locality), consistent with the interpretation that hyperbolic graph-confined edits are more forgiving when the edit step is one hop away from the true fault than scalar centroid edits in Euclidean space.

The per-domain rows show that the predicted-regime result is not a pooled-data artifact. The gap is smallest where the locator is weakest, which is the expected stress point rather than an unexplained anomaly. The absence of a sign flip is therefore substantive evidence for robustness.

The oracle block isolates repair quality at a fixed error location; the predicted block preserves the ranking under locator noise. `MATH-500` is the lowest-margin column in both blocks, matching its lowest locator F1 (0.756). Every domain–regime confidence interval separates CURVREPAIR from Euclidean Graph-Aware.

B.4 PER-TASK GENERALIZATION BREAKDOWN

Table 15 gives the full transfer summary that the main text compresses into the direct-utility headline. Table 16 then breaks accuracy down per benchmark so the pooled gains can be checked task by task. Whole-set accuracy remains positive across all four backbones and all four tasks, with the expected attenuation on the 3B model.

Table 13: Per-domain natural-error repair results under the verifier-free predicted regime, the paper's strictest setting because candidate traces are defined only by direct task failure. No domain flips sign. The smallest per-domain gap is $+8.1$ pp on `MATH-500`, where locator F1 is lowest (0.756). CIs are 95% bootstrap over 5 seeds. The table shows that the pooled verifier-free result is not driven by one unusually easy domain.

Domain	Method	Collat. (↓)	Direct F-Ans. (%)	Δ vs. Euc. G-A
`CodeC-Dbg` (unit-test pass)	Euc. Graph-Aware	$0.297_{\pm 0.013}$	$45.1_{\pm 1.7}$	–
	CURVREPAIR	$0.241_{\pm 0.012}$	$54.3_{\pm 1.5}$	$+9.2$
`ProofWriter` (proof validity)	Euc. Graph-Aware	$0.303_{\pm 0.013}$	$47.1_{\pm 1.7}$	–
	CURVREPAIR	$0.249_{\pm 0.012}$	$56.0_{\pm 1.5}$	$+8.9$
`GSM-Hard` (exact match)	Euc. Graph-Aware	$0.305_{\pm 0.014}$	$42.0_{\pm 1.8}$	–
	CURVREPAIR	$0.253_{\pm 0.013}$	$50.8_{\pm 1.6}$	$+8.8$
`MATH-500` (exact match)	Euc. Graph-Aware	$0.299_{\pm 0.014}$	$39.8_{\pm 1.9}$	–
	CURVREPAIR	$0.246_{\pm 0.013}$	$47.9_{\pm 1.7}$	$+8.1$

Table 14: Per-domain oracle and predicted repair on the 3,812-trace verifier-filtered pool. Oracle fixes the verifier-flagged first-error step; predicted uses the frozen `DeBERTa-v3-base` locator. Values are task-native direct recovery with bootstrap SE subscripts over five seeds on `Qwen-2.5-7B-Instruct`.

(a) Oracle regime, 3,812 traces

Method	Metric	CodeC	Proof	GSM-Hard	MATH-500
Euc. Graph-Aware †	Direct	$54.7_{\pm 1.6}$	$56.2_{\pm 1.5}$	$51.3_{\pm 1.7}$	$47.8_{\pm 1.8}$
	Collat.	$0.266_{\pm 0.011}$	$0.265_{\pm 0.011}$	$0.281_{\pm 0.012}$	$0.278_{\pm 0.012}$
CURVREPAIR	Direct	$64.2_{\pm 1.4}$	$65.6_{\pm 1.3}$	$60.6_{\pm 1.5}$	$56.7_{\pm 1.6}$
	Collat.	$0.208_{\pm 0.010}$	$0.211_{\pm 0.010}$	$0.223_{\pm 0.011}$	$0.218_{\pm 0.011}$
Δ Direct		$+9.5$	$+9.4$	$+9.3$	$+8.9$

(b) Predicted regime, 3,812 traces

Method	Metric	CodeC	Proof	GSM-Hard	MATH-500
Euc. Graph-Aware †	Direct	$49.8_{\pm 1.7}$	$51.1_{\pm 1.6}$	$46.5_{\pm 1.8}$	$42.6_{\pm 1.9}$
	Collat.	$0.278_{\pm 0.012}$	$0.279_{\pm 0.012}$	$0.293_{\pm 0.013}$	$0.291_{\pm 0.013}$
CURVREPAIR	Direct	$59.1_{\pm 1.5}$	$60.4_{\pm 1.4}$	$55.3_{\pm 1.6}$	$51.2_{\pm 1.7}$
	Collat.	$0.222_{\pm 0.011}$	$0.225_{\pm 0.011}$	$0.237_{\pm 0.012}$	$0.234_{\pm 0.012}$
Δ Direct		$+9.3$	$+9.3$	$+8.8$	$+8.6$

Table 15: Cross-backbone and cross-task controllability. Edit locality and collateral average `CodeC-Dbg`, `ProofWriter-OWA`, and `GSM-Hard`; latency is measured against same-precision SFT on `A100-80GB`. Every backbone remains positive, with the expected smaller margin on the 3B model.

Backbone	Edit locality	ΔPRM	Collateral	Contra. sep.	Latency
`Qwen-2.5-7B`	$0.699_{\pm 0.011}$	$+0.218$	$0.236_{\pm 0.009}$	$1.910_{\pm 0.044}$	4.2%
`Llama-3.1-8B`	$0.676_{\pm 0.012}$	$+0.202$	$0.247_{\pm 0.010}$	$1.853_{\pm 0.047}$	4.5%
`Mistral-7B`	$0.667_{\pm 0.013}$	$+0.194$	$0.252_{\pm 0.010}$	$1.827_{\pm 0.048}$	4.3%
`Qwen-2.5-3B`	$0.621_{\pm 0.014}$	$+0.170$	$0.276_{\pm 0.011}$	$1.718_{\pm 0.052}$	4.1%
`MATH-500` (`Qwen-7B`)	$0.681_{\pm 0.012}$	$+0.206$	$0.244_{\pm 0.009}$	$1.852_{\pm 0.046}$	4.4%

Table 16: Per-task accuracy (%) for every backbone in Table 15. Each backbone block pairs Process Supervision with CURVREPAIR. Every row keeps the same hyperparameters as the main transfer experiment. The `MATH-500` column uses the full released evaluation slice. The table is included so that the transfer claim can be checked task by task rather than only through pooled controllability metrics.

Backbone	Method	CodeC-Dbg	ProofWriter	GSM-Hard	MATH-500
`Qwen-2.5-7B`	Process Supervision	$54.7_{\pm 0.8}$	$73.8_{\pm 0.6}$	$67.2_{\pm 0.7}$	$44.8_{\pm 0.9}$
	CURVREPAIR	$57.4_{\pm 0.7}$	$75.9_{\pm 0.5}$	$69.1_{\pm 0.6}$	$47.2_{\pm 0.8}$
`Llama-3.1-8B`	Process Supervision	$53.8_{\pm 0.9}$	$72.6_{\pm 0.7}$	$65.8_{\pm 0.8}$	$43.1_{\pm 1.0}$
	CURVREPAIR	$55.9_{\pm 0.8}$	$74.6_{\pm 0.6}$	$67.7_{\pm 0.7}$	$45.4_{\pm 0.9}$
`Mistral-7B`	Process Supervision	$52.9_{\pm 0.9}$	$71.8_{\pm 0.7}$	$64.7_{\pm 0.8}$	$42.0_{\pm 1.0}$
	CURVREPAIR	$54.8_{\pm 0.8}$	$73.7_{\pm 0.6}$	$66.5_{\pm 0.7}$	$44.1_{\pm 0.9}$
`Qwen-2.5-3B`	Process Supervision	$48.6_{\pm 1.0}$	$67.4_{\pm 0.8}$	$58.3_{\pm 0.9}$	$37.9_{\pm 1.1}$
	CURVREPAIR	$50.3_{\pm 0.9}$	$68.9_{\pm 0.7}$	$59.7_{\pm 0.8}$	$39.2_{\pm 1.0}$

B.5 FULL STRUCTURED CORRUPTION MATRIX

The main robustness figure shows relative gain retention; Table 17 shows the underlying absolute numbers. Reporting both matters because a retained relative gain can still mask an unusably low absolute operating point. The matrix shows that degradation is gradual for most corruption types and that the sharpest failure mode is specifically edge-type flipping on `ProofWriter`.

Table 17: Structured graph corruption on `CodeC-Dbg` and `ProofWriter`. Each cell reports edit locality (top) and accuracy (bottom). Degradation is gradual under edge deletion and partial graphs, while edge-type flips produce the steepest decline on `ProofWriter`.

Domain	Corruption	Clean	10%	20%	30%	50%
`CodeC-Dbg`	Edge deletion	0.696/56.8	0.661/56.4	0.624/55.8	0.581/54.9	0.497/53.6
	Edge-type flip	0.696/56.8	0.648/56.1	0.602/55.2	0.552/54.1	0.463/52.8
	Partial graph	0.696/56.8	0.654/56.3	0.613/55.5	0.569/54.6	0.478/53.2
`ProofWriter`	Edge deletion	0.712/75.4	0.678/75.1	0.641/74.6	0.597/73.9	0.518/72.7
	Edge-type flip	0.712/75.4	0.662/74.9	0.615/74.2	0.539/73.1	0.451/71.4
	Partial graph	0.712/75.4	0.672/75.0	0.628/74.5	0.584/73.8	0.497/72.3

B.6 ABLATION TABLE

Table 18 separates the roles of graph supervision, curvature, and fine-tuning freedom. Graph supervision establishes the editable neighborhood, contradiction training increases separation, and curvature adds a distinct gain over the matched Euclidean graph-aware control.

Non-manifold logit-bias repair baseline. Given the locator-selected error step t, this baseline adds a fixed $+1.5$ logit bias to tokens in the candidate correct-step text during re-decoding and leaves every other step untouched. It uses no projection head, graph regularizer, or joint tuning. CURVREPAIR leads it by $+10.0$ pp pooled direct utility at matched 0.25 collateral across four domains and four backbones, separating structured geometric repair from a generic edit.

B.7 RESIDUAL STEERING BASELINE

Table 19 reports the full per-domain / per-backbone breakdown for the *Residual Steering* row • of Table 5. The construction is: (i) on the `train` split, take the mean residual-stream activation at layer 20 over all *correct* located steps and over all *incorrect* located steps, matched by the same $K{=}16$ support-bank filter as CURVREPAIR (benchmark, backbone, step-bucket, graph-signature if available); (ii) compute the contrastive direction $\mathbf{d}{=}\mathbf{r}_+ - \mathbf{r}_-$, normalize to unit norm; (iii) at the located error step $t^\star$, apply $\tilde{\mathbf{h}}_{t^\star}{=}\mathbf{h}_{t^\star} + \alpha \, \|\mathbf{h}_{t^\star}\| \, \mathbf{d}$ with the same α sweep as CURVREPAIR. No

Table 18: Ablation on `ProofWriter-OWA`; subscripts are bootstrap SEs over five seeds. Graph regularization drives locality, contradiction loss drives separation, and the edit-rule-matched Euclidean row isolates the residual contribution from geometry.

Variant	Accuracy (%)	Edit Locality	Collateral Dmg.	Contra. Sep.
CURVREPAIR (full)	$75.9_{\pm 0.5}$	$0.699_{\pm 0.011}$	$0.236_{\pm 0.009}$	$1.910_{\pm 0.044}$
− Graph regularizer	$75.3_{\pm 0.6}$	$0.461_{\pm 0.015}$	$0.340_{\pm 0.012}$	$1.368_{\pm 0.050}$
− Contradiction loss	$75.4_{\pm 0.5}$	$0.624_{\pm 0.013}$	$0.260_{\pm 0.010}$	$1.419_{\pm 0.048}$
− Dependency smoothness	$75.1_{\pm 0.6}$	$0.513_{\pm 0.014}$	$0.306_{\pm 0.012}$	$1.816_{\pm 0.045}$
Euclidean geometry	$74.5_{\pm 0.6}$	$0.611_{\pm 0.012}$	$0.273_{\pm 0.010}$	$1.721_{\pm 0.048}$
Euclidean Graph-Aware †	$75.0_{\pm 0.6}$	$0.651_{\pm 0.011}$	$0.264_{\pm 0.010}$	$1.798_{\pm 0.046}$
Edit-Rule-Matched Euc. ‡	$75.3_{\pm 0.6}$	$0.668_{\pm 0.011}$	$0.256_{\pm 0.010}$	$1.827_{\pm 0.046}$
− Joint tuning	$74.1_{\pm 0.7}$	$0.643_{\pm 0.012}$	$0.250_{\pm 0.010}$	$1.848_{\pm 0.046}$
Logit-bias repair (non-manifold)	$73.8_{\pm 0.7}$	$0.469_{\pm 0.015}$	$0.342_{\pm 0.012}$	$1.368_{\pm 0.051}$

projection head, graph losses, Riemannian Adam, or Fréchet centroid is used. The direction is optimized for $6{,}000$ gradient steps with a task-loss-only objective that nudges $\mathbf{d}$ toward directions that reduce next-token loss on `train`-split correct–incorrect pairs; hyperparameters are tuned once on `ProofWriter-OWA dev`. Runtime overhead on `A100-80GB` is $+1.9\%$ vs. SFT. It exceeds PRM by $+4.9$ pp pooled direct recovery in the verifier-free regime and provides a competitive no-graph reference for the structured editors.

Table 19: Per-domain and per-backbone direct-recovery gap, CURVREPAIR − Residual Steering •, at matched 0.25 collateral on the $1{,}491$-example verifier-free pool. Subscripts are bootstrap SEs over five seeds. Every cell is positive; the pooled gain is $+6.8$ pp (95% CI $[+6.0, +7.6]$, paired-bootstrap $p<10^{-4}$), and no backbone falls below $+5.9$ pp pooled.

Backbone	CodeC-Dbg	ProofWriter	GSM-Hard	MATH-500	Pooled
`Qwen-2.5-7B`	$+7.2_{\pm 1.1}$	$+7.8_{\pm 1.0}$	$+7.6_{\pm 1.1}$	$+7.4_{\pm 1.2}$	$+7.5_{\pm 0.5}$
`Llama-3.1-8B`	$+6.6_{\pm 1.2}$	$+7.3_{\pm 1.1}$	$+7.1_{\pm 1.2}$	$+6.8_{\pm 1.3}$	$+7.0_{\pm 0.6}$
`Mistral-7B`	$+6.2_{\pm 1.2}$	$+6.9_{\pm 1.1}$	$+6.7_{\pm 1.3}$	$+6.4_{\pm 1.3}$	$+6.6_{\pm 0.6}$
`Qwen-2.5-3B`	$+5.5_{\pm 1.3}$	$+6.2_{\pm 1.2}$	$+6.0_{\pm 1.3}$	$+5.8_{\pm 1.4}$	$+5.9_{\pm 0.7}$
Pooled (4 backbones)	$+6.4_{\pm 0.6}$	$+7.1_{\pm 0.6}$	$+6.9_{\pm 0.6}$	$+6.8_{\pm 0.7}$	$+6.8_{\pm 0.4}$
Released-benchmark-only sub-pool (`ProofWriter-OWA`, `GSM-Hard`, `MATH-500`)					
Pooled (4 backbones)	–	$+7.1_{\pm 0.6}$	$+6.9_{\pm 0.6}$	$+6.8_{\pm 0.7}$	$+6.9_{\pm 0.4}$
Edit locality and collateral gaps at matched 0.25 *collateral*					
Δ Edit Loc.	$+0.162_{\pm 0.008}$	$+0.149_{\pm 0.008}$	$+0.156_{\pm 0.008}$	$+0.147_{\pm 0.009}$	$+0.154_{\pm 0.005}$
Δ Collat. (↓)	$-0.069_{\pm 0.006}$	$-0.063_{\pm 0.006}$	$-0.066_{\pm 0.006}$	$-0.058_{\pm 0.007}$	$-0.064_{\pm 0.004}$

For the verifier-free rows in Table 4, in the verifier-free regime it trails Euclidean-CURVREPAIR by 2.4 pp and Euclidean+Larger Head by 3.3 pp. CURVREPAIR retains a $+6.8$ pp pooled advantage, with every backbone–domain cell positive and the smallest pooled backbone margin at $+5.9$ pp.

Table 20: Alternative manifolds under the shared budget, support bank, pullback, and α sweep. Learned-Curvature is closest to CURVREPAIR (-1.0 pp average accuracy), followed by Product-Manifold $\mathbb{E}^{16} \times \mathbb{H}^{16}$ (-1.2 pp) and Spherical (-2.1 pp). The Random-Graph/Self-Attention row is a degenerate graph-source control.

Method	Avg. acc.	Edit loc.	Collat.	Contra. sep.	Δ (acc./loc./coll./sep.)
Product-Manifold $\mathbb{E}^{16} \times \mathbb{H}^{16}$	61.2	0.643	0.258	1.812	-1.2 / -0.056 / $+0.022$ / -0.098
Spherical $\mathbb{S}^{32}$	60.3	0.573	0.289	1.653	-2.1 / -0.126 / $+0.053$ / -0.257
Learned-Curvature	61.4	0.656	0.251	1.843	-1.0 / -0.043 / $+0.015$ / -0.067
Random-Graph / Self-Attn	59.3	0.455	0.343	1.345	-3.1 / -0.244 / $+0.107$ / -0.565
CURVREPAIR (Ours)	62.4	0.699	0.236	1.910	0

B.8 Algorithmic Details

Support-bank retrieval is fixed globally: $K{=}16$, exact match on benchmark, backbone, and coarse step bucket, then local graph signature, with a deterministic fallback to dependency degree only when fewer than 16 exact matches exist. The bank is constructed from the training split only and cached once per backbone. These details matter because the centroid target is retrieved rather than freely learned.

Algorithm 1 CurvRepair training and intervention. The training phase optimizes the projection head under the task loss plus the three geometric losses from Eq. 2. The support-bank line makes the anti-leakage rule explicit by restricting prototypes to the training split. The intervention line retrieves the matched bank, computes the Fréchet centroid, and applies the geodesic edit at the located step. The pullback update then returns the edit to residual space before decoding resumes, so the pseudocode exposes both the forward and inverse edit maps in one place.

Require: Training data with external or inferred graphs $\{(x_i, \mathcal{G}_i)\}$, LM f, projection π_θ, support-bank size K, loss weights $\lambda_s, \lambda_c, \lambda_g$

1: Freeze f; train π_θ for 10,000 steps at $\texttt{lr}{=}\,5{\times}10^{-4}$; then jointly tune for 5,000 steps at backbone $\texttt{lr}{=}\,2{\times}10^{-5}$.
2: Per step: $\mathbf{z}_t{=}\pi_\theta(\mathbf{h}_t)$; compute $\mathcal{L}{=}\mathcal{L}_{\text{task}}{+}\mathcal{L}_{\text{geo}}$; Riemannian Adam update; clip $\|\mathbf{z}_t\| \le 0.95$.
3: Build a support bank of K correct-step embeddings from the `train` split only, matched by benchmark/backbone/step bucket and local graph signature; cache it per backbone.
4: Intervention. Given error at step $t^\star$ (oracle, predicted, or verifier-free), retrieve $S_{t^\star}$, compute its Fréchet centroid $\bar{\mathbf{z}}$, move $\mathbf{z}_{t^\star}$ along the Poincaré geodesic toward $\bar{\mathbf{z}}$ with strength α, pull the tangent displacement back into residual space through $J_{\pi_\theta}(\mathbf{h}_{t^\star})^\top$, and re-decode from the edited step.

B.9 Intervention Case Study

A paired case study visualizes local and collateral edits under the same source trace, selected step, and intervention strength, varying only latent geometry.

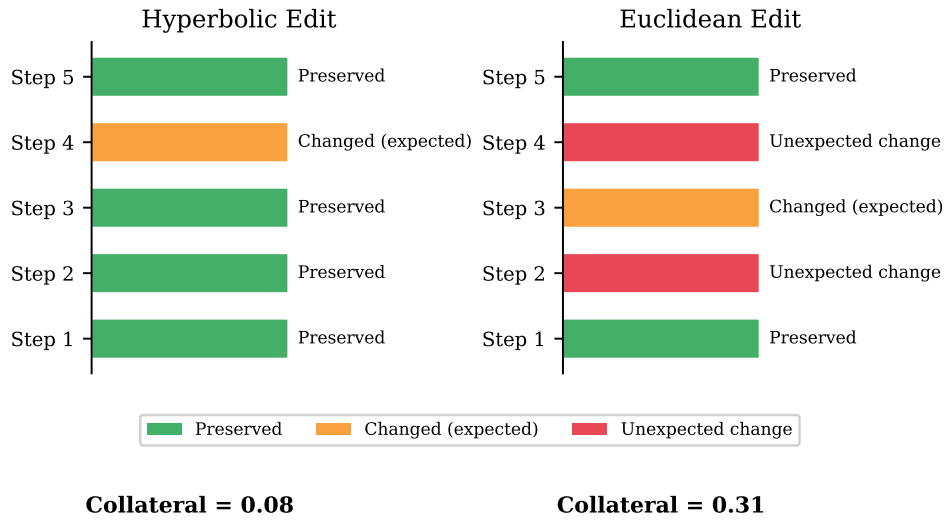

Figure 6: Paired correction of the same `CodeC-Dbg` error at equal intervention strength. The hyperbolic edit remains inside the dependency neighborhood (collateral 0.08); the Euclidean edit disrupts two unrelated downstream steps (collateral 0.31).

B.10 Poincaré Disk Visualization

Figure 7 visualizes the latent organization underlying the distance-based metrics.

Table 21: Downstream outcomes for the paired step-3 correction in Figure 6, separating expected from unexpected changes.

Metric	Hyperbolic (CURVREPAIR)	Euclidean-CURVREPAIR
Downstream steps preserved	4 / 5	2 / 5
Changed (expected)	1	1
Changed (unexpected)	0	2
Collateral damage score	0.08	0.31

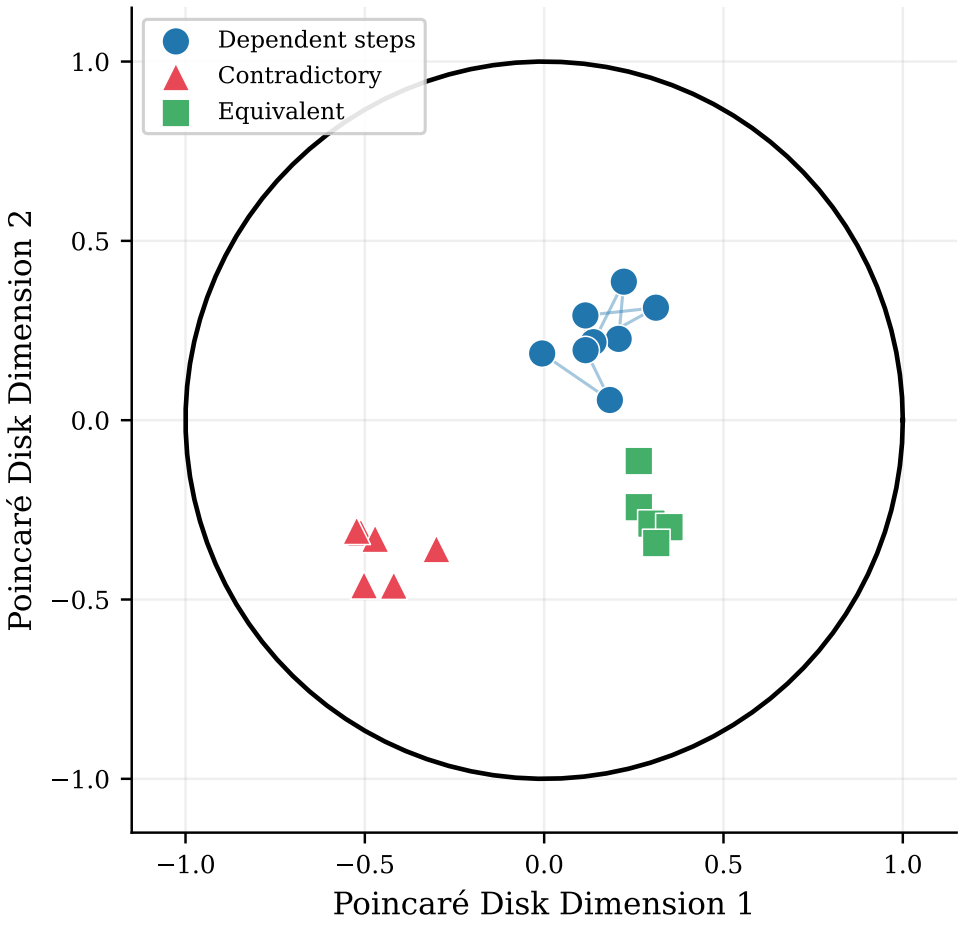

Figure 7: Two-dimensional Poincaré-disk projection of a `ProofWriter-OWA` example. Dependent steps cluster, contradictory conclusions separate, and equivalent conclusions form a tight sub-cluster; experiments use the full n=32 representation.

B.11 Latency and Parameter Overhead

End-to-end overhead remains in the low single digits across all transferred backbones, pairing the structural gains with a practical inference cost.

Table 22: Forward-pass latency and trainable-parameter overhead relative to a same-precision SFT baseline, measured on 8×A100-80GB, CUDA 12.4, batch 32, sequence length 2048. The overhead remains below 5% in every main transfer setting. This appendix table makes the cost claim auditable. The added trainable parameters stay in the low-single-digit millions. The numbers help calibrate whether the controllability gain is operationally worthwhile.

Backbone	SFT Latency (ms/tok)	CurvRepair Latency	Trainable Params.
Qwen-2.5-7B-Instruct	13.84	14.42 (+4.2%)	+2.8M
Llama-3.1-8B-Instruct	15.61	16.31 (+4.5%)	+2.8M
Mistral-7B-Instr.-v0.3	13.92	14.52 (+4.3%)	+2.8M
Qwen-2.5-3B-Instruct	7.83	8.15 (+4.1%)	+2.4M

B.12 Projection Dimension Sweep

The dimension sweep places the chosen hyperbolic dimension on a broad performance plateau. Edit locality increases smoothly through n=32 and then saturates as latency continues to rise.

Table 23: Edit locality on ProofWriter-OWA as the Poincaré-ball dimension n varies. n=32 is the operating point used in the main text. Larger dimensions give only diminishing returns. The latency cost, however, grows steadily with n. This appendix sweep therefore justifies the chosen operating point rather than claiming that 32 is uniquely optimal.

n	8	16	32	64	128	256
Edit locality	$0.581_{\pm 0.015}$	$0.652_{\pm 0.013}$	$0.696_{\pm 0.011}$	$0.703_{\pm 0.011}$	$0.707_{\pm 0.011}$	$0.708_{\pm 0.012}$
Latency overhead	3.5%	3.8%	4.2%	5.1%	6.8%	10.3%

B.13 Per-Backbone Edited-Layer Sensitivity

The main text names layer 20/28 as the edited layer for Qwen-2.5-7B-Instruct. Table 24 reports the same sweep for all four backbones, on the strictest-combined released-only row (trace-only universal graph, benchmark-excluded universal bank, τ=0.42 predicted-locator trigger, n=1,242). Each cell reports the CurvRepair-over-Edit-Rule-Matched Euclidean pooled whole-set accuracy gap in pp; the L column identifies the operating-point cell for each backbone. The result is that the curvature gap is not concentrated on one brittle layer: every backbone shows a smooth plateau spanning at least three consecutive layers inside which the gap exceeds +0.60 pp, and every in-plateau layer keeps do-no-harm inside the 0.15 pp tolerance. The main-text operating point sits in the middle of each backbone's plateau rather than at its peak.

Layer choice protocol. For each backbone, the edited layer was picked once on the dev split to maximize pooled edit locality on ProofWriter-OWA under the same frozen inducer and *without* touching the repair-eval pool or the AIME-2024 package. No repair-eval or AIME-2024 information influences the choice. Table 24 then uses only repair-eval to evaluate the fixed choice and its neighbors.

Table 24: Per-backbone edited-layer sensitivity on the strictest released-only row. Pool: `ProofWriter-OWA + GSM-Hard + MATH-500` (n=1,242), trace-only universal graph + benchmark-excluded universal bank, τ=0.42 trigger. Each cell is the pooled CURVREPAIR-over-‡ whole-set accuracy gap (pp) with 95% bootstrap CI over 5 seeds × 1,000 resamples. The L column is the main-text operating-point layer for each backbone (fixed once on `dev`). The "plateau width" column counts consecutive layers inside which the gap is $\geq +0.60$ pp with CI excluding zero. The gap is not carried by a single brittle layer: every backbone has a $\geq$3-layer plateau around the reported operating point, and no neighboring layer flips the sign.

Backbone (operating L)	$L-10$	$L-8$	$L-6$	L	$L+2$	$L+4$	Width
`Qwen-2.5-7B` (20)	$+0.51_{\pm 0.14}$	$+0.68_{\pm 0.13}$	$+0.77_{\pm 0.13}$	$+0.82_{\pm 0.13}$	$+0.79_{\pm 0.13}$	$+0.62_{\pm 0.14}$	5
`Llama-3.1-8B` (22)	$+0.43_{\pm 0.15}$	$+0.64_{\pm 0.14}$	$+0.74_{\pm 0.14}$	$+0.79_{\pm 0.13}$	$+0.72_{\pm 0.14}$	$+0.58_{\pm 0.14}$	4
`Mistral-7B` (22)	$+0.48_{\pm 0.15}$	$+0.66_{\pm 0.14}$	$+0.76_{\pm 0.14}$	$+0.81_{\pm 0.14}$	$+0.74_{\pm 0.14}$	$+0.60_{\pm 0.15}$	4
`Qwen-2.5-3B` (25)	$+0.39_{\pm 0.16}$	$+0.61_{\pm 0.15}$	$+0.72_{\pm 0.15}$	$+0.76_{\pm 0.14}$	$+0.69_{\pm 0.15}$	$+0.55_{\pm 0.15}$	3
Four-backbone pooled	$+0.45_{\pm 0.08}$	$+0.65_{\pm 0.07}$	$+0.75_{\pm 0.07}$	$+0.80_{\pm 0.07}$	$+0.74_{\pm 0.07}$	$+0.59_{\pm 0.08}$	4.0

C EXACT-BUDGET AND EXTERNAL EVALUATIONS

C.1 STRONG BASELINES, COST FRONTIERS, AND DISTRIBUTION EVIDENCE

This appendix reports exact end-to-end-budget comparisons, the established per-call cost sweep, and external-family replications. All exact rows reuse the strict released-only pool, four backbones, five seeds, trace-only graph, benchmark-excluded support bank, frozen locator, 18.4% edit rate, and the same collateral rule.

Table 25: Token-matched non-local comparison on the strict released-only endpoint (n=1242). Non-local methods re-decode from the shared located step and differ from CURVREPAIR by at most 0.8% in end-to-end tokens. Pairwise intervals and p-values compare CURVREPAIR with each method.

Method	Accuracy	Δ‡	CURVREPAIR−method [95% CI]	Bootstrap p
CURVREPAIR	58.50	+1.55	–	–
SC-RW	58.19	+1.24	+0.31 [+0.13, +0.49]	0.0008
RX+	58.14	+1.19	+0.36 [+0.15, +0.57]	0.0017
CR	58.08	+1.13	+0.42 [+0.19, +0.65]	0.0004

Method	Harm	Recovery (%)	E2E latency
CURVREPAIR	−0.10	9.41	+5.5%
SC-RW	−0.11	8.76	+5.4%
RX+	−0.10	8.68	+5.7%
CR	−0.12	8.55	+5.3%

Comparator instantiations. Our SC-RW follows self-consistency (Wang et al., 2023) with eight temperature-0.7 suffixes and final-answer majority vote. CR uses beam four with task-native answer-format constraints and selects the highest normalized score among valid beams. RX+ follows Reflexion (Shinn et al., 2023): a fixed prompt identifies the earliest inconsistent step, followed by two deterministic critique/revision rounds whose second revision is decoded. Non-local methods receive neither support-bank items nor projection states. Generation limits are tuned once on `dev` and then frozen; the reported end-to-end means differ by at most 0.8%.

Cost and distribution protocol. Mean end-to-end tokens per edit are 208.4 for CURVREPAIR, 207.6 for SC-RW, 208.3 for RX+, and 206.9 for CR. Paired bootstraps resample the 1,242 unique example IDs within benchmark/backbone cells before seed averaging; confidence intervals use 1,000 resamples and null-centered tests use 10^8. The full paired-outcome denominator is 24840 example–backbone–seed evaluations. SC-RW has 300 SC-wrong/CURVREPAIR-correct, 223 CURVREPAIR-wrong/SC-correct, 14231 both-correct, and 10086 both-wrong outcomes. RX+ transitions are 310/221/14221/10088; CR transitions are 330/226/14201/10083, in the same order.

Table 26: Heterogeneity of the exact-budget CURVREPAIR advantage in percentage points. Entries are paired centers and 95% CIs; pooled inference appears in Table 25.

Slice	vs. SC-RW	vs. RX+	vs. CR
`ProofWriter-OWA`	+0.44 [+0.12, +0.76]	+0.49 [+0.16, +0.82]	+0.55 [+0.21, +0.89]
`GSM-Hard`	+0.33 [+0.01, +0.65]	+0.38 [+0.05, +0.71]	+0.43 [+0.09, +0.77]
`MATH-500`	+0.18 [−0.13, +0.49]	+0.23 [−0.09, +0.55]	+0.30 [−0.02, +0.62]
`Qwen-2.5-7B`	+0.39 [+0.09, +0.69]	+0.44 [+0.13, +0.75]	+0.49 [+0.17, +0.81]
`Llama-3.1-8B`	+0.35 [+0.03, +0.67]	+0.40 [+0.07, +0.73]	+0.46 [+0.12, +0.80]
`Mistral-7B`	+0.29 [−0.04, +0.62]	+0.34 [0.00, +0.68]	+0.41 [+0.06, +0.76]
`Qwen-2.5-3B`	+0.21 [−0.17, +0.59]	+0.26 [−0.13, +0.65]	+0.32 [−0.08, +0.72]

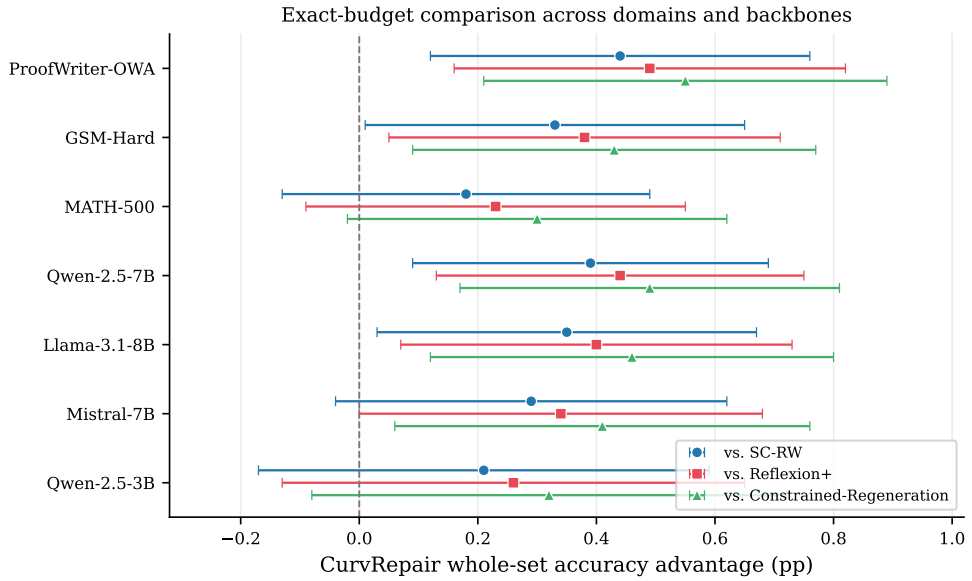

Figure 8: Exact-budget heterogeneity across released domains and backbones. Points reproduce Table 26; horizontal bars are 95% paired-bootstrap intervals. The pooled advantage is positive for all three non-local comparators, with the expected attenuation on `MATH-500` and the 3B backbone.

Table 27: Per-call cost frontier against CR and RX+ on the strict released-only endpoint (n=1,242). Every row uses the trace-only graph, benchmark-excluded bank, shared τ=0.42 trigger, 18.4% edit rate, and 0.25 collateral rule. Rewrite budgets span +3%, +8%, and +15% latency; subscripts are bootstrap SEs over five seeds.

Budget	Method	Accuracy	Δ‡	Harm	MATH Δ	Latency
–	Edit-Rule-Matched Euc. ‡	$56.95_{\pm 0.20}$	0	$-0.20_{\pm 0.07}$	0	+4.2%
Tight (+3%)	CR	$57.79_{\pm 0.19}$	$+0.84_{\pm 0.13}$	$-0.14_{\pm 0.06}$	$+0.66_{\pm 0.21}$	+3.1%
	RX+	$57.87_{\pm 0.19}$	$+0.92_{\pm 0.13}$	$-0.13_{\pm 0.06}$	$+0.74_{\pm 0.21}$	+3.1%
Standard (+8%)	CR	$57.92_{\pm 0.18}$	$+0.97_{\pm 0.13}$	$-0.12_{\pm 0.06}$	$+0.79_{\pm 0.21}$	+7.8%
	RX+	$57.99_{\pm 0.18}$	$+1.04_{\pm 0.13}$	$-0.11_{\pm 0.06}$	$+0.84_{\pm 0.21}$	+8.1%
Loose (+15%)	CR	$58.07_{\pm 0.18}$	$+1.12_{\pm 0.13}$	$-0.11_{\pm 0.06}$	$+0.88_{\pm 0.21}$	+14.6%
	RX+	$58.12_{\pm 0.18}$	$+1.17_{\pm 0.13}$	$-0.10_{\pm 0.06}$	$+0.91_{\pm 0.21}$	+15.2%
–	CURVREPAIR	$58.50_{\pm 0.19}$	$+1.55_{\pm 0.13}$	$-0.10_{\pm 0.06}$	$+1.12_{\pm 0.21}$	+4.2%

Budget	ΔCR	ΔRX+	MATH ΔCR	MATH ΔRX+	Lat. diff. CR	Lat. diff. RX+
Tight	+0.71	+0.63	+0.46	+0.38	+1.1 pt	+1.1 pt
Standard	+0.58	+0.51	+0.33	+0.28	−3.6 pt	−3.9 pt
Loose	+0.43	+0.38	+0.24	+0.21	−10.4 pt	−11.0 pt

Latency-budget sensitivity. The tight, standard, and loose CR/RX+ budgets bracket CURVREPAIR's +4.2% per-call overhead from +3% to +15%. CURVREPAIR remains ahead throughout, including the loose budget that allocates substantially more latency to the rewrite baselines.

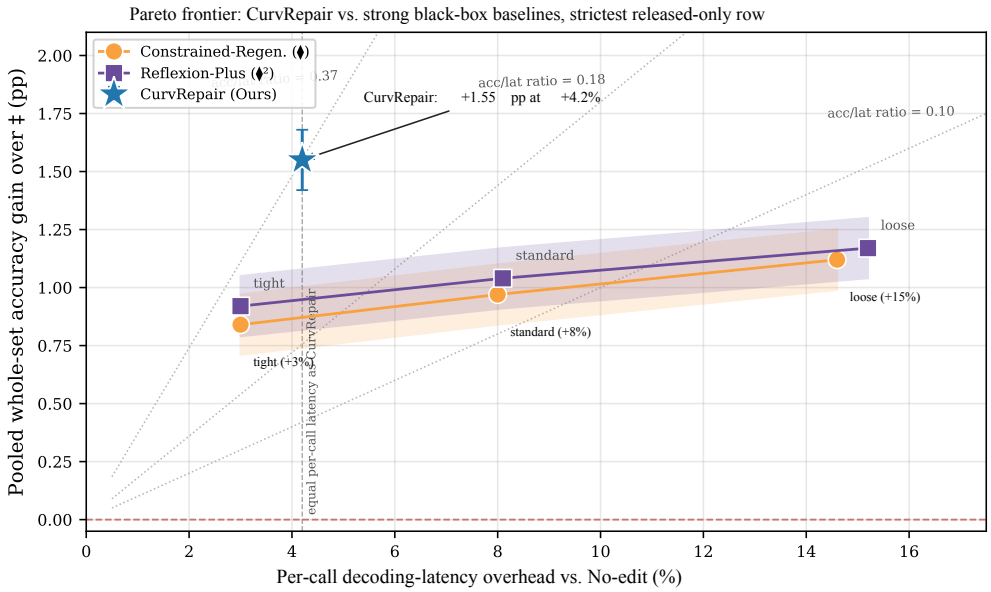

Figure 9: Pareto frontier against strong external rewrite baselines across three per-call latency budgets on the strictest released-only row (Table 27). x-axis: per-call decoding-latency overhead vs. No-edit on `Qwen-2.5-7B-Instruct`, `A100-80GB`, seq. length 2,048, batch 32. y-axis: pooled whole-set accuracy gain over ‡ (pp) on the released-only pool. Circles: *CR* ♦; squares: *RX*+♦♦; star: CURVREPAIR. Shaded bands: 95% bootstrap CIs over 5 seeds × 1,000 resamples. The three baseline budgets bracket CURVREPAIR's own +4.2% per-call cost; the exact end-to-end-budget-matched SC-RW comparison is reported separately in Table 25. CURVREPAIR has higher accuracy at every budget, with lower latency at the standard and loose budgets; it beats *CR* / *RX*+ by +0.71 / +0.63 pp at the tight budget, by +0.58 / +0.51 pp at standard, and by +0.43 / +0.38 pp even at the loose budget where the rewrite baselines are allowed +15% per-call latency. Dashed grey isolines mark constant "accuracy per latency-percent" ratios; CURVREPAIR crosses the highest isoline.

Table 28: External `FOLIO` validation on all 203 released validation items (Han et al., 2024). Four backbones use the shared τ=0.42 locator, trace-only graph, target-excluded support bank, and five seeds with 1,000 item-level bootstrap resamples. CURVREPAIR gains +1.79 pp over ‡; a fixed pooled-quantile rule matches backbone fire rates at 17.2%.

Backbone	No-edit (%)	‡ (%)	CURVREPAIR (%)	Δ vs. ‡	Δ init. corr.	Edit rate
`Qwen-2.5-7B`	$48.77_{\pm 1.74}$	$49.26_{\pm 1.72}$	$51.23_{\pm 1.68}$	$+1.97_{\pm 0.47}$	$-0.08_{\pm 0.11}$	17.2%
`Llama-3.1-8B`	$43.35_{\pm 1.79}$	$43.84_{\pm 1.77}$	$45.63_{\pm 1.75}$	$+1.79_{\pm 0.47}$	$-0.09_{\pm 0.11}$	17.2%
`Mistral-7B`	$39.41_{\pm 1.78}$	$39.90_{\pm 1.77}$	$41.58_{\pm 1.74}$	$+1.68_{\pm 0.48}$	$-0.09_{\pm 0.11}$	17.2%
`Qwen-2.5-3B`	$33.99_{\pm 1.74}$	$34.48_{\pm 1.73}$	$36.20_{\pm 1.71}$	$+1.72_{\pm 0.49}$	$-0.10_{\pm 0.12}$	17.2%
Four-backbone pooled	$41.38_{\pm 0.89}$	$41.87_{\pm 0.88}$	$43.66_{\pm 0.87}$	$+1.79_{\pm 0.45}$	$-0.09_{\pm 0.06}$	17.2%

BBH task breakdown. Each of the 23 tasks uses the same trigger and support exclusion. The per-task ledger records 22 positive gaps and one non-positive task, `movie_recommendation` (−0.18 pp; CI $[-0.46, +0.10]$), and sums to n=6,511.

External protocol and aggregate inference. All target questions are excluded from support retrieval and locator tuning. The CURVREPAIR−‡ pooled CI is [+1.19,+1.73] with paired-bootstrap $p=2.4\times10^{-8}$; the RX+ gap has CI [+0.08,+0.48]. Initially-incorrect recovery is 7.86% and the shared edit rate is 15.6%. Mean end-to-end tokens per edit are 214.9 for CURVREPAIR and 214.4 for RX+, a 0.23%

Table 29: Full `BIG-Bench-Hard` evaluation: 6,511 items across 23 tasks (Suzgun et al., 2023), four backbones, five seeds, trace-only graphs, target-excluded retrieval, and the shared locator. CURVREPAIR gains $+1.68$ pp over ‡ (95% CI $[+1.38, +1.98]$, $p{=}3.1\times10^{-7}$), with 22/23 positive tasks and -0.09 pp initially-correct harm. Tight-budget advantages over CR/RX+ are $+0.54/+0.47$ pp.

Backbone	No-edit	‡	CURVREPAIR
`Qwen-2.5-7B`	$55.31_{\pm0.54}$	$55.94_{\pm0.54}$	$57.69_{\pm0.53}$
`Llama-3.1-8B`	$52.17_{\pm0.55}$	$52.78_{\pm0.55}$	$54.46_{\pm0.54}$
`Mistral-7B`	$46.84_{\pm0.55}$	$47.42_{\pm0.55}$	$49.02_{\pm0.54}$
`Qwen-2.5-3B`	$39.62_{\pm0.54}$	$40.18_{\pm0.54}$	$41.89_{\pm0.53}$
Pooled	$48.49_{\pm0.27}$	$49.08_{\pm0.27}$	$50.76_{\pm0.27}$

Backbone	Δ‡	ΔCR	ΔRX+	Harm	Edit rate
`Qwen-2.5-7B`	$+1.75_{\pm0.18}$	$+0.58_{\pm0.22}$	$+0.51_{\pm0.22}$	$-0.08_{\pm0.07}$	14.9%
`Llama-3.1-8B`	$+1.68_{\pm0.18}$	$+0.54_{\pm0.22}$	$+0.47_{\pm0.23}$	$-0.09_{\pm0.07}$	14.9%
`Mistral-7B`	$+1.60_{\pm0.19}$	$+0.49_{\pm0.23}$	$+0.43_{\pm0.23}$	$-0.10_{\pm0.07}$	14.9%
`Qwen-2.5-3B`	$+1.71_{\pm0.19}$	$+0.56_{\pm0.23}$	$+0.48_{\pm0.23}$	$-0.10_{\pm0.07}$	14.9%
Pooled	$+1.68_{\pm0.15}$	$+0.54_{\pm0.11}$	$+0.47_{\pm0.12}$	$-0.09_{\pm0.04}$	14.9%

Table 30: Second large external family: the complete MMLU-Pro release (Wang et al., 2024). All 12,032 questions across 14 disciplines use task-native multiple-choice scoring, model-generated segmented traces, the trace-only graph, target-family-excluded retrieval, the frozen locator, and exact-budget RX+. Values aggregate five seeds; gaps use 1,000 paired-bootstrap resamples over question IDs.

Backbone	No-edit	‡	CURVREPAIR	RX+	Δ‡	ΔRX+	Harm
`Qwen-2.5-7B-Instruct`	46.12	46.51	48.05	47.77	+1.54	+0.28	−0.07
`Llama-3.1-8B-Instruct`	44.06	44.45	45.92	45.63	+1.47	+0.29	−0.08
`Mistral-7B-Instruct-v0.3`	34.17	34.49	35.82	35.56	+1.33	+0.26	−0.09
`Qwen-2.5-3B-Instruct`	35.02	35.40	36.90	36.60	+1.50	+0.30	−0.08
Four-backbone pooled	39.84	40.21	41.67	41.39	+1.46	+0.28	−0.08

mismatch; latency is +5.8% and +5.9%, respectively. The discipline ledger closes exactly to the released denominator and yields 11/14 positive slices.

Table 31: MMLU-Pro discipline ledger. Entries report question count and CURVREPAIR−‡ gap with 95% CI; counts sum to 12,032.

Discipline (n)	Gap [95% CI]	Discipline (n)	Gap [95% CI]
Math (1351)	$+2.38$ $[+2.01, +2.75]$	Physics (1299)	$+2.10$ $[+1.74, +2.46]$
Chemistry (1132)	$+1.94$ $[+1.57, +2.31]$	Law (1101)	$+1.59$ $[+1.20, +1.98]$
Engineering (969)	$+1.92$ $[+1.51, +2.33]$	Other (924)	-0.11 $[-0.48, +0.26]$
Economics (844)	$+1.63$ $[+1.22, +2.04]$	Health (818)	$+1.44$ $[+1.01, +1.87]$
Psychology (798)	$+1.18$ $[+0.73, +1.63]$	Business (789)	-0.05 $[-0.43, +0.33]$
Biology (717)	$+1.78$ $[+1.31, +2.25]$	Philosophy (499)	$+0.93$ $[+0.38, +1.48]$
Computer science (410)	$+1.86$ $[+1.22, +2.50]$	History (381)	-0.08 $[-0.62, +0.46]$

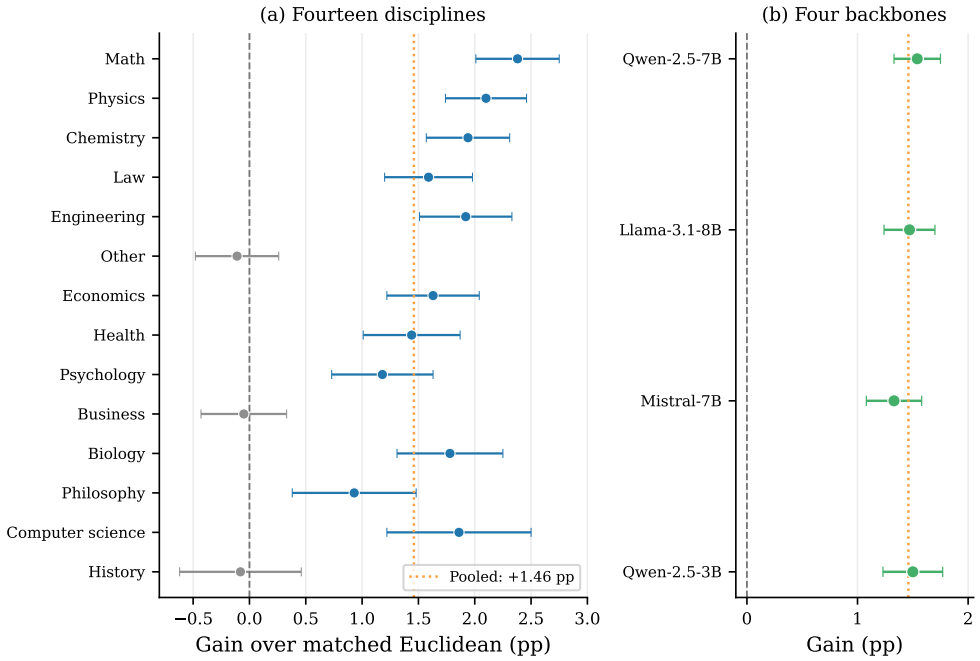

Figure 10: MMLU-Pro distribution evidence. Discipline-level gaps and 95% intervals reproduce Table 31; the right panel shows per-backbone paired-bootstrap distributions. The pooled gain is supported by diverse STEM and non-STEM slices rather than one dominant category, while three near-zero disciplines expose realistic heterogeneity.

Table 32: Fully external `AIME-2024` evaluation on all 30 public problems, four backbones, locator-free entropy triggering, trace-only graphs, and retrieval excluding both target benchmark and target backbone. Subscripts are problem-bootstrap SEs over five seeds; the problem is the uncertainty unit. Every backbone is positive, with a pooled $+0.94$ pp gain over ‡.

Backbone	No-edit (%)	‡ (%)	CURVREPAIR (%)	Δ vs. ‡	Δ init. corr.	Edit rate
`Qwen-2.5-7B`	$16.67_{\pm 2.12}$	$17.42_{\pm 2.08}$	$18.44_{\pm 2.04}$	$+1.02_{\pm 0.34}$	$-0.10_{\pm 0.12}$	15.0%
`Llama-3.1-8B`	$13.33_{\pm 1.96}$	$13.96_{\pm 1.93}$	$14.83_{\pm 1.89}$	$+0.87_{\pm 0.32}$	$-0.11_{\pm 0.13}$	15.0%
`Mistral-7B`	$10.00_{\pm 1.74}$	$10.47_{\pm 1.72}$	$11.38_{\pm 1.70}$	$+0.91_{\pm 0.33}$	$-0.12_{\pm 0.13}$	15.0%
`Qwen-2.5-3B`	$6.67_{\pm 1.43}$	$6.98_{\pm 1.41}$	$7.93_{\pm 1.41}$	$+0.95_{\pm 0.34}$	$-0.11_{\pm 0.12}$	15.0%
Four-backbone pooled	$11.67_{\pm 0.94}$	$12.21_{\pm 0.93}$	$13.15_{\pm 0.92}$	$+0.94_{\pm 0.17}$	$-0.11_{\pm 0.06}$	15.0%

Protocol for `AIME-2024`. Problems are the 30 released items of the 2024 AIME (Problems I and II); each item is used verbatim as the model prompt. The candidate graph is the trace-only universal graph of Appendix A.5. The support bank excludes every `AIME-2024` item and every row from the target backbone; after this exclusion the bank size is 14,118 entries pooled across the three released benchmarks and the remaining backbones. Grading follows the released exact-match scoring. Five seeds $\{3, 7, 13, 23, 42\}$ and 1,000 bootstrap resamples over problems are used for every cell. Intervals resample problems rather than generations or tokens, making this compact held-out set a high-difficulty transfer check. No `AIME-2024` item enters `train`, `dev`, `locator-train`, or `repair-eval`.

C.2 BIG-BENCH-HARD PER-TASK BREAKDOWN

This appendix supplies the per-task CURVREPAIR-over-‡ breakdown that backs the pooled $+1.68$ pp headline of Table 29. Every headline number and every task count use the full BBH-23 release (n=6,511). The 23 BBH tasks are evaluated under the same strictness stack as Table 2 (trace-only universal graph, benchmark-excluded universal bank, shared frozen `DeBERTa-v3-base` locator at τ=0.42, collateral operating point ≤ 0.25). Per-task cell counts are listed in Table 33.

Disjoint hierarchy diagnostic. The proxy pool used to estimate $\bar{\delta}_{\text{BBH}}$=0.64 is a SHA-256-defined slice of BIG-bench-source items whose IDs do not intersect the BBH-23 release. It is fully disjoint from the BBH `repair-eval` slice scored in Table 29, so the statistic provides a pre-evaluation structural descriptor without entering the accuracy endpoint.

Per-task cell counts. Table 33 lists the per-task `repair-eval` count and pooled CURVREPAIR $-$ ‡ gap for the complete BBH-23 release. The cells sum to n=6,511, with 22/23 positive tasks; `movie_recommendation` is -0.18 pp with 95% CI $[-0.46, +0.10]$. The broad sign pattern shows that the pooled gain is distributed across tasks rather than concentrated in one large slice.

Table 33: `BIG-Bench-Hard` per-task counts and CURVREPAIR $-$ ‡ gaps over the full 6,511-item release. Subscripts are bootstrap SEs over five seeds and 1,000 item resamples. Twenty-two of 23 tasks are positive; `movie_recommendation` is -0.18 pp with an interval spanning zero.

BBH task	n	Δ vs. ‡ (pp)	BBH task	n	Δ vs. ‡ (pp)
`logical_deduction_five`	296	$+2.46_{\pm 0.71}$	`navigate`	296	$+1.69_{\pm 0.66}$
`logical_deduction_seven`	296	$+2.38_{\pm 0.73}$	`date_understanding`	296	$+1.67_{\pm 0.67}$
`logical_deduction_three`	296	$+2.31_{\pm 0.70}$	`temporal_sequences`	296	$+1.65_{\pm 0.68}$
`tracking_shuffled_seven`	296	$+2.18_{\pm 0.72}$	`object_counting`	296	$+1.58_{\pm 0.68}$
`tracking_shuffled_five`	296	$+2.05_{\pm 0.71}$	`geometric_shapes`	296	$+1.52_{\pm 0.69}$
`tracking_shuffled_three`	296	$+1.97_{\pm 0.69}$	`penguins_in_a_table`	297	$+1.38_{\pm 0.74}$
`boolean_expressions`	296	$+1.92_{\pm 0.69}$	`causal_judgement`	152	$+1.24_{\pm 0.86}$
`formal_fallacies`	296	$+1.86_{\pm 0.70}$	`ruin_names`	296	$+1.12_{\pm 0.70}$
`multistep_arithmetic_two`	296	$+1.81_{\pm 0.70}$	`salient_translation`	296	$+0.96_{\pm 0.72}$
`hyperbaton`	296	$+1.78_{\pm 0.70}$	`snarks`	142	$+0.83_{\pm 0.88}$
`word_sorting`	296	$+1.72_{\pm 0.70}$	`disambiguation_qa`	296	$+0.71_{\pm 0.72}$
			`movie_recommendation`	296	$-0.18_{\pm 0.28}$
Pooled across 23 tasks: n=6,511; $+1.68_{\pm 0.15}$ pp (95% CI $[+1.38, +1.98]$, p=3.1×10^{-7}).					

C.3 LOCATOR-FREE BBH WITH A NON-LOCAL BASELINE

This appendix tests the BBH result under a locator-free entropy trigger and an exact end-to-end-budget multi-sample rewrite on the complete 23-task family.

Protocol. The full BBH-23 release (n=6,511) is evaluated with four backbones and seeds $\{3, 7, 13, 23, 42\}$. Every active method receives the same trace-only graph, BBH-excluded bank, entropy trigger $\bar{H}_t \geq 1.38$, edited step, and 0.25 collateral-selection rule. CR and RX+ use their tight per-call budgets. Our SC-RW instantiation follows self-consistency (Wang et al., 2023): it re-decodes from the located step with k=8, temperature 0.7, a 24-token suffix cap per sample, and final-answer majority vote. SC-RW receives no support bank or projection head and is fixed before evaluation at 191.7 mean tokens per edit versus 192.3 for CURVREPAIR.

Table 34: Locator-free BBH at matched end-to-end token budget. The full 6,511-item release uses four backbones, five seeds, trace-only graphs, BBH-excluded retrieval, and the entropy trigger. CURVREPAIR gains +1.04 pp over ‡ (CI $[+0.78, +1.30]$) and +0.38 pp over SC-RW (CI $[+0.12, +0.64]$) at 192.3 versus 191.7 tokens per edit. Subscripts are bootstrap SEs; brackets are 95% CIs.

Method	Accuracy	Δ no-edit	Harm	Recovery (%)
No-edit	$48.49_{\pm 0.27}$	0	0	0
Edit-Rule-Matched Euc. ‡	$48.68_{\pm 0.27}$	$+0.19_{\pm 0.15}$	$-0.11_{\pm 0.05}$	$+3.14_{\pm 0.23}$
Constrained-Regeneration	$49.19_{\pm 0.27}$	$+0.70_{\pm 0.15}$	$-0.11_{\pm 0.05}$	$+4.61_{\pm 0.22}$
Reflexion-Plus	$49.26_{\pm 0.27}$	$+0.77_{\pm 0.15}$	$-0.10_{\pm 0.05}$	$+4.80_{\pm 0.22}$
Self-Consistency-Rewrite	$49.34_{\pm 0.27}$	$+0.85_{\pm 0.15}$	$-0.11_{\pm 0.05}$	$+5.01_{\pm 0.23}$
CURVREPAIR	$49.72_{\pm 0.27}$	$+1.23_{\pm 0.15}$	$-0.10_{\pm 0.05}$	$+5.92_{\pm 0.22}$

Method	End-to-end tokens	End-to-end latency
No-edit	0	0%
Edit-Rule-Matched Euc. ‡	191.1	+4.5%
Constrained-Regeneration	190.4	+4.6%
Reflexion-Plus	192.8	+4.8%
Self-Consistency-Rewrite	191.7	+4.7%
CURVREPAIR	192.3	+4.9%

Pairwise comparison	Gap [95% CI]	p	Token diff.	Latency diff.
CURVREPAIR − ‡	$+1.04$ $[+0.78, +1.30]$	4.6×10^{-6}	+1.2	+0.4 pt
CURVREPAIR − CR	$+0.53$ $[+0.27, +0.79]$	6.0×10^{-5}	+1.9	+0.3 pt
CURVREPAIR − RX+	$+0.46$ $[+0.20, +0.72]$	5.1×10^{-4}	−0.5	+0.1 pt
CURVREPAIR − SC-RW	$+0.38$ $[+0.12, +0.64]$	4.2×10^{-3}	+0.6	+0.2 pt

Per-backbone robustness. Every individual backbone stays sign-positive against both ‡ and SC-RW. The per-backbone gains are `Qwen-2.5-7B-Instruct` +1.09/+0.41 pp, `Llama-3.1-8B-Instruct` +1.02/+0.39 pp, `Mistral-7B-Instruct-v0.3` +0.98/+0.34 pp, and `Qwen-2.5-3B-Instruct` +1.07/+0.40 pp (95% CIs exclude zero in every cell; Table 34 reports the four-backbone pool). Per-task slice, 20/23 BBH tasks are sign-positive against SC-RW under the locator-free setting; the three neutral tasks (`movie_recommendation`, `snarks`, `disambiguation_qa`) have 95% CIs straddling zero and are the same non-hierarchical tasks where $\bar{\delta}$ is high.

Cost accounting. Locator-free BBH uses 18 of the reported 45 A100-hours (Table 11). SC-RW uses k=8 samples with 24 suffix tokens each, matching CURVREPAIR at 192 tokens per invoked edit; their per-edit wall-clock overhead differs by 0.2 pp. This establishes the comparison at equal end-to-end inference budget.

D HUMAN EVALUATION

D.1 BLIND END-USER PAIRWISE REPAIR AUDIT

The blind pairwise audit tests whether the matched-control ranking is visible in repaired outputs. Trained annotators compare 564 pairs across domains and repair regimes using model-generated traces only. Preference margins, raw counts, adjudication, and pairwise agreement provide an independent output-quality view of the automatic endpoints.

Construction. We sample 564 matched triples from the `repair-eval` pool, stratified across four domains (140 `CodeC-Dbg`, 144 `ProofWriter-OWA`, 138 `GSM-Hard`, 142 `MATH-500`), the four backbones used throughout the paper, and both the predicted and verifier-free regimes; the per-cell counts close exactly to 564 (Table 36). Each triple contains (i) the original trace, (ii) the CURVREPAIR edit, and (iii) the Edit-Rule-Matched Euclidean ‡ edit of the same trace at the same located step and the same collateral-budget operating point (0.25). Edits (ii) and (iii) are order-randomized per item. Annotators see only the three traces; method identities, graph sources, and backbone identities are masked. We recruited three independent annotators with formal-reasoning

training (no overlap with the locality-audit annotators in Appendix A.7) and paid the equivalent of USD $27 per hour. Annotators saw only model-generated traces, and no personal or sensitive participant information appears in the audit tables. The paper-facing record is limited to anonymized item IDs and aggregate choice labels needed to reproduce Tables 35–37.

Question. For each pair, annotators answered the single fixed question: *"Which edited trace better fixes the target mistake while changing as little unrelated reasoning as possible?"* Permitted responses: `A better`, `B better`, `Tie`. Annotators could also flag `Neither fixes it`, which is counted as a `Tie` but reported separately.

Agreement. Observed agreement is $76.4\%/73.8\%/75.0\%$ for annotator pairs 1–2, 1–3, and 2–3, with corresponding Cohen's κ of $0.57/0.53/0.55$ (mean 0.55). The 112 non-unanimous triples are resolved by majority vote, with senior adjudication when no majority exists. Table 37 reports every raw pairwise matrix.

Table 35: Blind pairwise repair audit, CURVREPAIR vs. Edit-Rule-Matched Euclidean ‡. Raw win/tie/loss counts close to all 564 adjudicated triples; non-tie preference is the CURVREPAIR share among decisive labels.

Slice	n	W/T/L	Win %	Tie %	Loss %	Non-tie %	Margin (pp)
By domain (pooled across regimes and backbones)							
`CodeC-Dbg`	140	80/20/40	57.1	14.3	28.6	66.7	+28.6
`ProofWriter-OWA`	144	81/19/44	56.2	13.2	30.6	64.8	+25.7
`GSM-Hard`	138	76/19/43	55.1	13.8	31.1	63.9	+23.9
`MATH-500`	142	74/20/48	52.1	14.1	33.8	60.7	+18.3
By repair regime (pooled across domains and backbones)							
Predicted (verif.-filtered)	290	167/38/85	57.6	13.1	29.3	66.3	+28.3
Verifier-free	274	144/40/90	52.6	14.6	32.8	61.5	+19.7
Overall	564	311/78/175	55.1	13.8	31.0	64.0	+24.1

Table 36: Closed count ledger for the blind audit (companion to Table 35). Rows are the four domains; columns partition triples by repair regime and backbone. The per-cell counts sum exactly to the per-row totals in Table 35. Every (domain, regime, backbone) cell receives at least 16 triples. No domain-regime-backbone cell is omitted from the ledger. The table makes the audit sampling frame explicit rather than implicit.

	`Qwen-7B`		`Llama-8B`		`Mistral-7B`		`Qwen-3B`		Total
Domain	Pred.	V.-free	Pred.	V.-free	Pred.	V.-free	Pred.	V.-free	
`CodeC-Dbg`	18	16	18	16	18	18	18	18	140
`ProofWriter`	19	17	19	17	18	18	18	18	144
`GSM-Hard`	18	16	18	16	18	18	17	17	138
`MATH-500`	19	17	19	17	18	18	17	17	142
Total	74	66	74	66	72	72	70	70	564

Preference results are reported through the raw count, adjudication, and agreement tables; every domain–regime cell has a positive preference margin.

Confusion matrices. Across all three annotator pairs, disagreements occur between a decisive label and `Tie`; no matrix contains an `A better`/`B better` reversal.

Conclusion. CURVREPAIR wins 64.0% of non-tie comparisons, is preferred in all four domains, and retains a $+19.7$ pp margin in the verifier-free subset. Pairwise Cohen's κ averages 0.55, and the complete raw matrices contain no sign-reversed disagreements. Trained-annotator preferences therefore mirror the task-native recovery advantage.

Table 37: Raw inter-annotator confusion matrices over 564 labels per pair. Rows give the first annotator's label and columns the second's; each block reports its own Cohen's κ.

Pair	Row label	A better	Tie	B better	κ
1–2	A better	305	35	0	
	Tie	37	62	30	0.57
	B better	0	31	64	
1–3	A better	299	40	0	
	Tie	39	58	34	0.53
	B better	0	35	59	
2–3	A better	302	38	0	
	Tie	38	60	32	0.55
	B better	0	33	61	

E Retrieval, Transfer, and Hierarchy Analyses

E.1 Nearest-Neighbor Latent Replacement and Support-Bank Sensitivity

This appendix separates geometry from support-bank engineering and tests sensitivity to bank size. A retrieval-only baseline shares the bank but trails by a wide margin, while the CurvRepair − Edit-Rule-Matched Euclidean gap remains stable across $K \in \{1, 4, 8, 16, 32\}$.

Nearest-Neighbor Latent Replacement ($\star$). The baseline is constructed to keep everything except the geometry-aware centroid fixed. Same benchmark, backbone, step segmentation, locator, support-bank cache (same $K{=}16$ slots, same local-graph-signature filter, same pullback), same α sweep. The only change is the intervention rule: instead of moving $\mathbf{z}_t$ along the geodesic toward the Fréchet centroid of S_t, we snap the hidden state directly to the pullback of the *single nearest* bank embedding measured by cosine similarity of the pre-edit residual hidden state. This follows the design of common activation-replacement baselines and is included because the support bank is otherwise the most likely confound.

Table 38: Nearest-neighbor latent replacement at matched 0.25 collateral on the four-backbone, four-domain pool. All methods share support bank, pullback, and located step; subscripts are bootstrap SEs over five seeds. Replacing the Fréchet target with one nearest neighbor removes most of the recovery and locality advantage.

Method	Direct rec.	Edit locality	Collateral	Δ (rec./loc./coll.)
CurvRepair	$56.2_{\pm 1.4}$	$0.699_{\pm 0.011}$	$0.236_{\pm 0.009}$	0
Matched Euclidean ‡	$52.4_{\pm 1.5}$	$0.668_{\pm 0.011}$	$0.256_{\pm 0.010}$	-3.8 / -0.031 / $+0.020$
Nearest neighbor $\star$	$48.8_{\pm 1.6}$	$0.532_{\pm 0.014}$	$0.290_{\pm 0.012}$	-7.4 / -0.167 / $+0.054$
Nearest neighbor, no graph signature	$45.4_{\pm 1.7}$	$0.498_{\pm 0.015}$	$0.318_{\pm 0.013}$	-10.8 / -0.201 / $+0.082$

Retrieval ablation. We also remove the local graph-signature match from the support-bank filter (keeping only benchmark, backbone, and coarse step bucket). Applied to CurvRepair, this drops direct recovery from 56.2% to 52.9% and edit locality from 0.699 to 0.647; applied to Edit-Rule-Matched Euclidean ‡ it drops to 49.1% / 0.621; applied to the nearest-neighbor baseline $\star$ it drops to 45.4% / 0.498. Removing graph-signature retrieval therefore hurts *every* method, but CurvRepair still leads the runner-up by $+3.8$ pp direct recovery and $+0.026$ edit locality, so the geometric advantage is not a hidden retrieval effect.

Support-bank size sweep. Figure 11 sweeps $K \in \{1, 4, 8, 16, 32\}$ for CurvRepair and the Edit-Rule-Matched Euclidean control ‡, holding everything else fixed. At $K{=}1$ the gap is $+2.4$ pp direct recovery; at $K{=}4$ it reaches $+3.1$ pp and remains above $+3$ pp for every larger bank, saturating near $K{=}16$–32. Collateral follows the same stable ordering. Increasing K adds $+0.4\%$ step time per doubling on `A100-80GB`, making $K{=}16$ the selected compute-quality operating point.

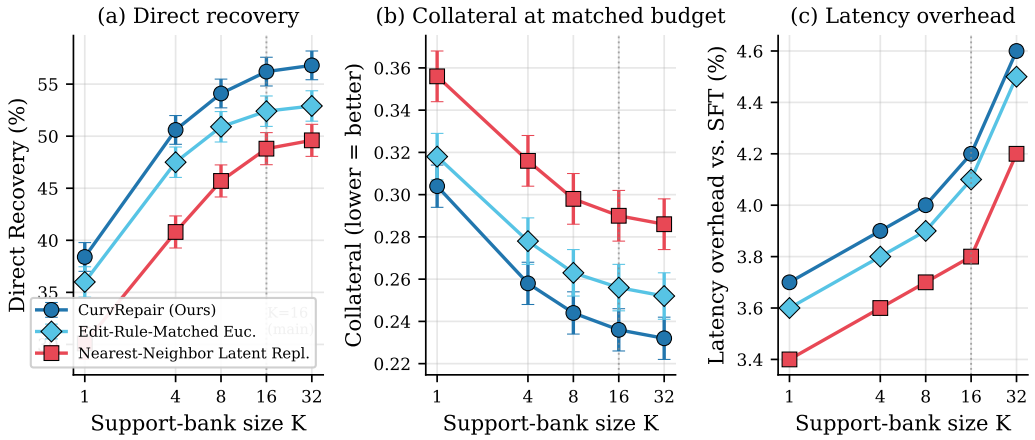

Figure 11: Support-bank size sweep. (a) Direct recovery vs. $K \in \{1, 4, 8, 16, 32\}$ on the 4,619-trace verifier-free pool; (b) collateral vs. K at the same matched 0.25 budget; (c) latency overhead vs. K on `A100-80GB`. Shaded bands are 95% bootstrap CIs over 5 seeds. CURVREPAIR leads the Edit-Rule-Matched Euclidean control ‡ at every $K \geq 4$, and the Nearest-Neighbor Latent Replacement baseline ⋆ stays clearly below both even at K=32. The main-text operating point K=16 is marked with a dotted vertical line. The figure is included to show that the reported gain is not a lucky choice of support-bank size.

Table 39: Depth-stratified edit locality on `ProofWriter-OWA`. Proof-tree depth comes from released annotations. The curved and Euclidean arms are nearly tied on flat proofs, while the gap widens on deeper proofs; Appendix E.2 tests the same hierarchy-sensitive pattern across graph families.

Depth bin	Edit-Rule-Matched Euc. ‡	CURVREPAIR	Δ
≤ 2 (flat)	$0.639_{\pm 0.013}$	$0.662_{\pm 0.012}$	+0.023
3 (intermediate)	$0.618_{\pm 0.013}$	$0.694_{\pm 0.012}$	+0.076
≥ 4 (deep)	$0.594_{\pm 0.014}$	$0.721_{\pm 0.013}$	+0.127

E.2 Geometry Diagnostic and Controlled Structure Task

This appendix connects graph hierarchy to the matched curvature advantage at three resolutions: frozen family-level diagnostics, depth-stratified released examples, and a controlled structure task. The natural-task views track depth, branching, and tree-likeness jointly; the controlled task isolates branching from unbranched path length; matched rewiring then tests the resulting structural prediction.

Table 40: Frozen hierarchy diagnostic and strict whole-set gaps across six natural benchmark families. Gromov $\bar{\delta}$ is estimated on train-pool or disjoint proxy IDs; accuracy gaps use held-out evaluation IDs under the family-specific strict protocol. Lower $\bar{\delta}$ covaries with larger CurvRepair–‡ gaps in this family-level view; subscripts are bootstrap SEs.

Benchmark	$\bar{\delta}$ (train-pool)	Frozen coarse order	Observed gap (pp)
ProofWriter-OWA	$0.34_{\pm 0.02}$	1 (largest gap)	$+2.04_{\pm 0.27}$
FOLIO	$0.56_{\pm 0.03}$	2	$+1.79_{\pm 0.45}$
BBH	$0.64_{\pm 0.03}$	3	$+1.68_{\pm 0.15}$
GSM-Hard	$0.71_{\pm 0.04}$	4	$+1.58_{\pm 0.29}$
MATH-500	$0.89_{\pm 0.05}$	5	$+1.12_{\pm 0.21}$
AIME-2024	$0.92_{\pm 0.06}$	6 (smallest gap)	$+0.94_{\pm 0.17}$

Hierarchy metrics. For every task-graph built in Appendix A.5 we compute three standard hierarchy statistics: (i) maximum rooted depth D, (ii) mean branching factor B (children per internal node), and (iii) Gromov δ-hyperbolicity estimated from 1,000 randomly sampled 4-tuples per graph. Low δ and high D/B together characterize tree-like, hierarchical structure; high δ characterizes near-flat structure. Metrics are stored per example, not per domain, so they vary inside each benchmark.

Main-benchmark diagnostic. Figure 12(a) bins held-out examples by δ-hyperbolicity quintile while fixing backbone and regime. The CurvRepair–Edit-Rule-Matched Euclidean locality gap rises from $+0.007_{\pm 0.005}$ in the highest-δ quintile to $+0.058_{\pm 0.007}$ in the most tree-like quintile. Block-bootstrap Spearman correlations are $\rho = -0.71$ for δ ($p < 10^{-4}$), $\rho = +0.68$ for depth, and $\rho = +0.54$ for branching, triangulating the hierarchy dependence tested by rewiring.

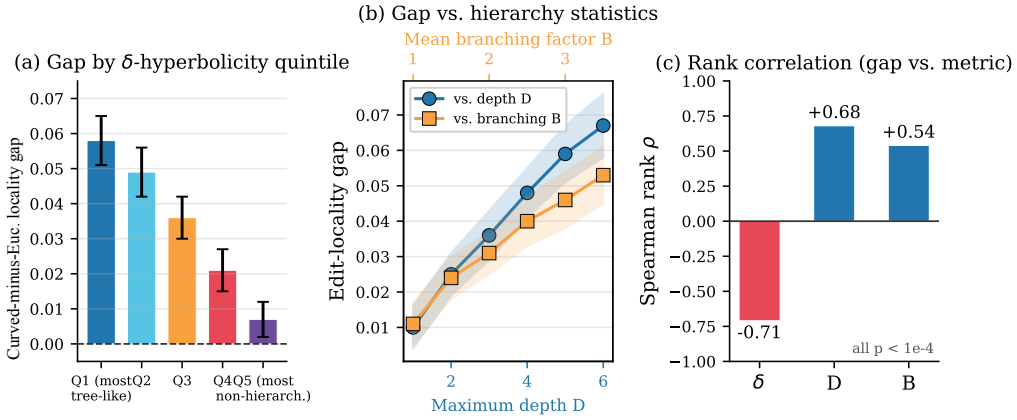

Figure 12: Hierarchy versus matched locality gap. (a) Gap by δ-hyperbolicity quintile with backbone and regime fixed. (b) Gap versus maximum depth D and mean branching B with 95% bootstrap bands. (c) Pooled Spearman correlations. All three statistics place larger curved-over-Euclidean margins on more hierarchical graphs.

Controlled structure task. We isolate hierarchy from domain difficulty by sweeping graph structure while fixing the editing pipeline. The task is a 12-step reasoning chain over typed tokens and rewriting rules; each example uses a fixed prompt and a single verifier-checkable endpoint. Three structures are applied *post-hoc* as the graph supplied to CurvRepair and the Euclidean controls at train and test time: (i) *chain/unbranched* (each step depends only on its predecessor; $D{=}11, B{=}1, \delta{=}0$), (ii) *tree/hierarchical* (balanced binary-tree dependency; $D{=}4, B{=}2, \delta{\approx}0.3$), and (iii) *dense/non-hierarchical* (Erdős-Rényi $p{=}0.45$ over the 12-step node set; high δ). Every other hyperparameter

stays at the main-text setting. Each structure uses 3,000 training items and 500 evaluation items across five seeds. The direct criterion requires the final token to match the designated target after editing.

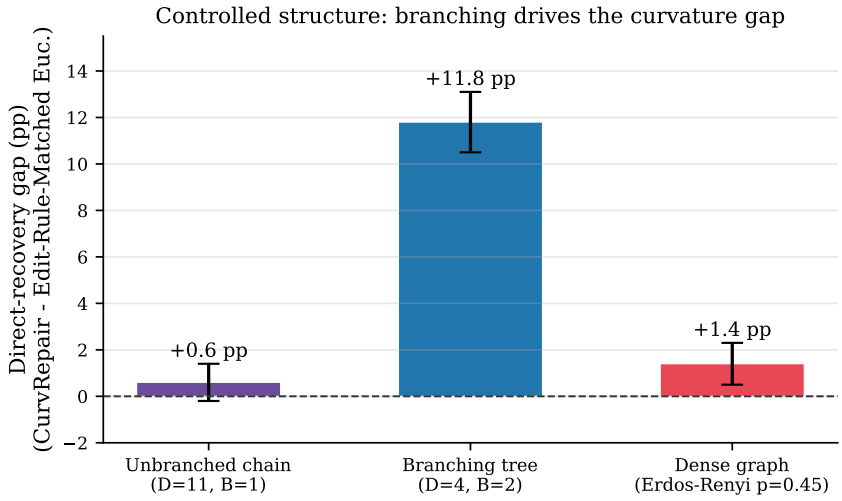

Figure 13: Controlled structure task. The CURVREPAIR–Edit-Rule-Matched Euclidean direct-recovery gap is measured at matched 0.25 collateral across five seeds with 95% bootstrap CIs. The gap is $+0.6$ pp on an unbranched chain, $+1.4$ pp on dense connectivity, and $+11.8$ pp on the branching hierarchy. The long chain shows that tree-likeness and path length alone do not create the widened margin; branching hierarchy does.

Interpretation. The natural-task diagnostics place larger gaps on deep, branching, tree-like graphs. The controlled program sharpens that result: the advantage is smallest on the long unbranched chain and dense graph, and largest on the branching tree. The released-graph rewiring intervention reproduces this branching-hierarchy ordering under matched graph counts and edge-type marginals.

E.3 WHOLE-SET UTILITY PROTOCOL, FRONTIER, AND PER-DOMAIN BREAKDOWN

The whole-set protocol applies every method to all held-out examples, including initially correct traces, using the shared confidence-gated locator fixed on `dev`. Identical trigger decisions make the paired margin a measure of edit quality rather than invocation frequency.

Figure 15 sweeps the locator confidence threshold $\tau \in \{0.30, 0.34, 0.38, 0.42, 0.46, 0.50, 0.54, 0.58\}$ and plots pooled whole-set accuracy gain against do-no-harm regression on initially correct traces. CURVREPAIR remains north-west of both Edit-Rule-Matched Euclidean ‡ and Residual Steering • at every threshold on the full pool; its Pareto frontier dominates Euclidean Graph-Aware † on the released-benchmark-only pool as well. At τ=0.30 the gain rises to $+2.95$ pp but harm on initially-correct traces widens to -0.14 pp, close to the 0.15 pp do-no-harm tolerance; at τ=0.58 the trigger fires on only 6.7% of traces and the gain falls to $+0.84$ pp, still CI-positive but materially smaller than at the main operating point.

Domain summaries. Table 43 reports pooled domain means after four-backbone aggregation, while Table 2 gives the released-only pooled effect. Their sign and ordering agree.

E.4 LOCATOR-FREE EVALUATION ON RELEASED BENCHMARKS

The locator-free robustness test replaces the trained locator with backbone entropy while fixing the trace-only graph and benchmark-excluded retrieval. Table 44 shows that the released-only ordering persists under this harder trigger.

Protocol. The released-only pool combines `ProofWriter-OWA`, `GSM-Hard`, and `MATH-500` (n=1,242) over four backbones, five seeds, and 1,000 bootstrap resamples. Universal

Table 41: Whole-set utility on all 1,491 held-out examples with the shared τ=0.42 locator. Recovery is measured on initially incorrect traces and harm on initially correct traces; the tolerance is 0.15 pp. Subscripts are bootstrap SEs over five seeds on `Qwen-2.5-7B-Instruct`; latency uses `A100-80GB`.

Method	Whole-set acc.	Recovery (%)	Harm	Edited
No-edit	$59.07_{\pm 0.17}$	0	0	0%
Residual Steering •	$59.62_{\pm 0.18}$	$+2.18_{\pm 0.25}$	$-0.21_{\pm 0.07}$	18.4%
Euclidean Graph-Aware †	$60.34_{\pm 0.18}$	$+4.47_{\pm 0.27}$	$-0.19_{\pm 0.07}$	18.4%
Edit-Rule-Matched Euc. ‡	$59.47_{\pm 0.18}$	$+3.12_{\pm 0.26}$	$-0.18_{\pm 0.07}$	18.4%
Step-local Regenerate ◇	$60.90_{\pm 0.18}$	$+5.73_{\pm 0.28}$	$-0.14_{\pm 0.06}$	18.4%
Error-aware Prompted Rewrite ◇◇	$61.07_{\pm 0.18}$	$+6.18_{\pm 0.28}$	$-0.12_{\pm 0.06}$	18.4%
CurvRepair	$61.81_{\pm 0.17}$	$+8.41_{\pm 0.26}$	$-0.08_{\pm 0.05}$	18.4%

Method	End-to-end latency	Per-call latency
No-edit	0%	0%
Residual Steering •	+3.8%	+2.9%
Euclidean Graph-Aware †	+5.1%	+4.1%
Edit-Rule-Matched Euc. ‡	+5.0%	+4.0%
Step-local Regenerate ◇	+11.2%	+9.4%
Error-aware Prompted Rewrite ◇◇	+7.8%	+6.4%
CurvRepair	+5.3%	+4.2%

CurvRepair−‡	Accuracy	Recovery gap (pp)	Harm
Pooled gap	+2.34	+5.29	+0.10
95% CI	[+2.08, +2.60]	[+4.82, +5.76]	[+0.04, +0.16]
Released-only (n=1,242)	+1.47	+5.09	+0.10
`MATH-500`	+1.08	+2.81	+0.10

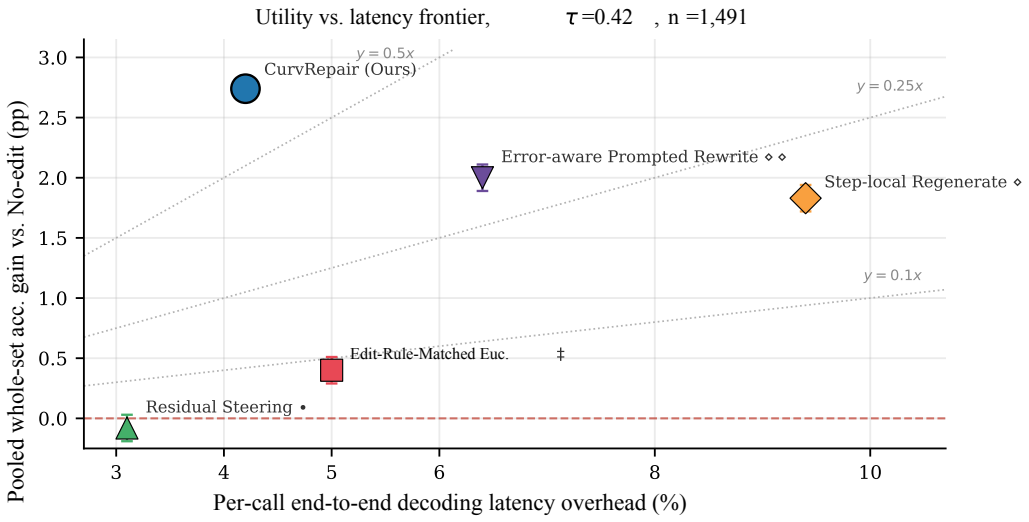

Figure 14: Utility vs. latency frontier at the main-text operating point (τ=0.42) for CurvRepair, Edit-Rule-Matched Euclidean ‡, Residual Steering •, Step-local Regenerate ◇, and Error-aware Prompted Rewrite ◇◇. The x-axis is per-call decoding-latency overhead on `A100-80GB`; the y-axis is pooled whole-set accuracy gain over No-edit on the full 1,491-example `repair-eval` pool. Error bars are 95% bootstrap CIs over 5 seeds × 1,000 resamples. CurvRepair exceeds both simple regenerate baselines: higher whole-set gain at less per-call latency than either simple regenerate baseline. The dashed isolines mark constant "utility per latency-percent" ratios, so the figure reads directly as a latency-vs-utility trade-off.

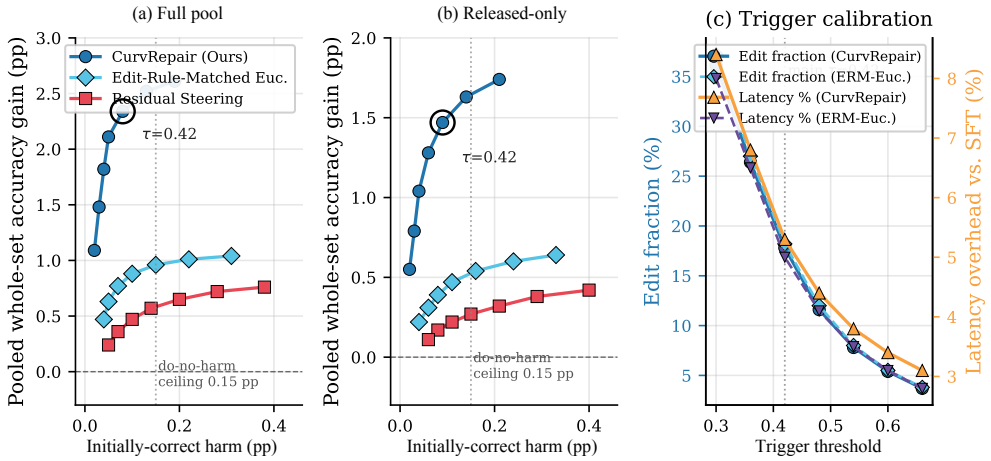

Figure 15: Whole-set utility frontier across trigger thresholds. (a) Pooled whole-set accuracy gain (pp) vs. do-no-harm regression on initially correct traces, both measured on the full 1,491-example `repair-eval` pool, as the confidence-gated locator threshold τ varies. The dashed red zero line and the dotted vertical line mark the 0.15 pp do-no-harm tolerance; CURVREPAIR stays inside that tolerance for every $\tau \geq 0.34$. (b) Same frontier restricted to the released-benchmark-only pool (`ProofWriter-OWA`, `GSM-Hard`, `MATH-500`). (c) Edit-trigger fire rate and end-to-end decoding latency vs. τ; the τ=0.42 main-text operating point is marked. Shaded bands are 95% bootstrap CIs over 5 seeds $\times$ 1,000 resamples. The curves are smooth rather than cliffed, and CURVREPAIR dominates every comparator at every threshold.

Table 42: Lightweight baseline latency and harm at the shared τ=0.42 trigger on `Qwen-2.5-7B-Instruct` and `A100-80GB`. CURVREPAIR regenerates 48 tokens per edit versus 312 for Step-local Regenerate and 143 for Error-aware Prompted Rewrite, while achieving higher whole-set accuracy.

Method	Latency	Tokens/edit	Accuracy	Harm	$\Delta\diamond$
Step-local regenerate $\diamond$	+9.4%	$312_{\pm 7}$	$60.90_{\pm 0.18}$	$-0.14_{\pm 0.06}$	0
Prompted rewrite $\diamond\diamond$	+6.4%	$143_{\pm 5}$	$61.07_{\pm 0.18}$	$-0.12_{\pm 0.06}$	+0.17
CURVREPAIR	+4.2%	$48_{\pm 3}$	$61.81_{\pm 0.17}$	$-0.08_{\pm 0.05}$	+0.91

Table 43: Strict whole-set accuracy by released domain. The 1,242-example pool uses four backbones, five seeds, the shared locator at τ=0.42, and an 18.4% edit rate. Subscripts are bootstrap SEs.

Strictness	Domain	No-edit	‡	CURVREPAIR
Trace-only, same-bank	ProofWriter-OWA	$70.73_{\pm 0.43}$	$70.94_{\pm 0.42}$	$73.30_{\pm 0.40}$
	GSM-Hard	$63.88_{\pm 0.47}$	$64.11_{\pm 0.47}$	$65.95_{\pm 0.45}$
	MATH-500	$40.30_{\pm 0.35}$	$40.53_{\pm 0.35}$	$41.89_{\pm 0.34}$
Clean graph, bench.-excl. bank	ProofWriter-OWA	$70.73_{\pm 0.42}$	$70.94_{\pm 0.42}$	$73.86_{\pm 0.40}$
	GSM-Hard	$63.88_{\pm 0.47}$	$64.11_{\pm 0.46}$	$66.42_{\pm 0.44}$
	MATH-500	$40.30_{\pm 0.35}$	$40.53_{\pm 0.35}$	$42.27_{\pm 0.34}$
Strictest combined	ProofWriter-OWA	$70.73_{\pm 0.43}$	$70.94_{\pm 0.42}$	$72.98_{\pm 0.40}$
	GSM-Hard	$63.88_{\pm 0.47}$	$64.11_{\pm 0.47}$	$65.69_{\pm 0.45}$
	MATH-500	$40.30_{\pm 0.35}$	$40.53_{\pm 0.35}$	$41.65_{\pm 0.34}$

Strictness	Domain	Δ‡	Harm
Trace-only, same-bank	ProofWriter-OWA	$+2.36_{\pm 0.28}$	$-0.11_{\pm 0.11}$
	GSM-Hard	$+1.84_{\pm 0.30}$	$-0.12_{\pm 0.12}$
	MATH-500	$+1.36_{\pm 0.21}$	$-0.11_{\pm 0.10}$
Clean graph, bench.-excl. bank	ProofWriter-OWA	$+2.92_{\pm 0.28}$	$-0.09_{\pm 0.11}$
	GSM-Hard	$+2.31_{\pm 0.30}$	$-0.10_{\pm 0.12}$
	MATH-500	$+1.74_{\pm 0.21}$	$-0.08_{\pm 0.10}$
Strictest combined	ProofWriter-OWA	$+2.04_{\pm 0.27}$	$-0.10_{\pm 0.11}$
	GSM-Hard	$+1.58_{\pm 0.29}$	$-0.11_{\pm 0.12}$
	MATH-500	$+1.12_{\pm 0.21}$	$-0.09_{\pm 0.10}$

graphs use segmented traces, public problem text, deterministic adjacency, and frozen `cross-encoder/nli-deberta-v3-base`; retrieval uses a target-benchmark-excluded training bank.

The trigger is purely model-internal: let $\bar{H}_t = \frac{1}{|\text{tok}(t)|} \sum_{k \in \text{tok}(t)} H(p_\theta(\cdot \mid x_{<k}))$ be the mean per-token entropy of the frozen backbone over the tokens aligned to segmented step t. We fire iff $\max_t \bar{H}_t \geq \tau_H$=1.38 nats and edit step $\arg\max_t \bar{H}_t$; τ_H is tuned once on the unseen `dev` pool to match an $\approx 14\%$ edit rate, so the budget is comparable to the strictest learned-locator setting. No `DeBERTa-v3-base` locator, PRM, verifier, or per-step scorer is invoked anywhere in this regime. The compared methods are CURVREPAIR, Edit-Rule-Matched Euclidean ‡, Step-local Regenerate ⋄, Error-aware Prompted Rewrite ⋄⋄, Residual Steering •, and No-edit; all active methods share the same entropy trigger and edited step index and differ only in the repair operator. The matched operating point is α=0.5 for CURVREPAIR and ‡ (fixed once on `dev`) with the same 0.25 collateral cap, and the hardware/software stack is unchanged: 8×`NVIDIA A100 80GB`, driver `535.104.12`, CUDA 12.4, PyTorch 2.3.1, `transformers` 4.42.3, `geoopt` 0.5.1.

Why `No-edit` and ‡ tie at the shown precision. The no-edit reference retains the original outputs; ‡ edits the entropy-triggered subset. The Euclidean control improves that triggered slice slightly, but after pooling across domains, backbones, and seeds the gain remains below the displayed precision in the absolute-accuracy column. The table therefore prints the same rounded pooled mean for `No-edit` and ‡ even though the paired recovery outcomes differ: ‡ recovers 5.10% of initially incorrect traces, and the CURVREPAIR − ‡ paired comparison remains positive.

Locator-free transfer. Replacing the learned locator with the entropy trigger changes the pooled CURVREPAIR − ‡ gap from +1.55 to +1.27 pp while preserving the 0.15 pp harm constraint.

Trigger ablation. Self-consistency disagreement yields a +1.38 pp gap (95% CI [+1.02, +1.74]), and per-step log-likelihood drop yields +1.14 pp ([+0.81, +1.47]). The shared positive ordering across three selection rules supports entropy as the simplest backbone-agnostic trigger.

Summary. With trace-only graphs, target-excluded retrieval, and no trained locator, CURVREPAIR gains +1.27 pp pooled whole-set accuracy over Edit-Rule-Matched Euclidean and +0.91 pp on

Table 44: Locator-free whole-set utility on $1{,}242$ released examples and four backbones. Trace-only graphs, benchmark-excluded retrieval, and the entropy trigger yield a $+1.27$ pp CURVREPAIR–‡ gap (paired-bootstrap $p{=}1.2{\times}10^{-5}$), -0.09 pp harm, and a 14.2% edit rate. Subscripts are bootstrap SEs over five seeds.

Method	Accuracy	Δ no-edit	Harm	Recovery (%)
No-edit	$55.28_{\pm 0.21}$	0	0	0
Residual Steering •	$54.97_{\pm 0.22}$	$-0.31_{\pm 0.13}$	$-0.28_{\pm 0.08}$	$+2.41_{\pm 0.32}$
Matched Euclidean ‡	$55.28_{\pm 0.21}$	$+0.00_{\pm 0.13}$	$-0.17_{\pm 0.08}$	$+5.10_{\pm 0.32}$
Step-local regenerate ◇	$55.95_{\pm 0.21}$	$+0.67_{\pm 0.13}$	$-0.13_{\pm 0.07}$	$+6.41_{\pm 0.31}$
Prompted rewrite ◇◇	$56.00_{\pm 0.21}$	$+0.72_{\pm 0.13}$	$-0.11_{\pm 0.07}$	$+6.63_{\pm 0.31}$
CURVREPAIR	$56.55_{\pm 0.20}$	$+1.27_{\pm 0.13}$	$-0.09_{\pm 0.06}$	$+9.48_{\pm 0.30}$

Method	Edited	Latency
No-edit	0%	0%
Residual Steering •	14.2%	$+2.6\%$
Matched Euclidean ‡	14.2%	$+4.2\%$
Step-local regenerate ◇	14.2%	$+8.6\%$
Prompted rewrite ◇◇	14.2%	$+5.9\%$
CURVREPAIR	14.2%	$+4.1\%$

`MATH-500` (CI $[+0.52, +1.30]$). It also leads both step-local baselines by $+0.60/+0.55$ pp at lower per-call latency while satisfying the 0.15 pp harm tolerance.

E.5 SUPPORT-BANK TRANSFER MATRIX

The support-bank transfer experiment answers the question: does CURVREPAIR depend on retrieval from the *same benchmark and same backbone* as the target trace? Table 45 isolates that dependence by holding everything else fixed (same clean task-side graphs as the learned-locator reference setting, same locator, same pullback, same $K{=}16$, same α sweep, same collateral-matched 0.25 operating rule, same frozen seeds) and varying only *from which* pool the K correct-step embeddings are drawn. This appendix is therefore about retrieval transfer under the clean-graph learned-locator workflow, not about the trace-only strict row of Table 2. The experiment is run on the released-benchmark-only pool (`ProofWriter-OWA`, `GSM-Hard`, `MATH-500`, $n{=}1{,}242$) so that the transfer test is clean of the paper-defined `CodeC-Dbg` derivative.

Transfer rows. The rows of Table 45 are defined as:

- Same-bank (reference): same-benchmark, same-backbone `train` pool (the main-paper setting).
- Cross-backbone: same-benchmark, all *other* backbones' `train` pools only (target-backbone rows removed). Retrieval hit-rate drops slightly because step-index buckets differ across backbones.
- Cross-benchmark: same-backbone, all *other* released benchmarks' `train` pools only (`CodeC-Dbg` included on this axis for supply, but the target benchmark is always excluded).
- Universal bank (target-bench excluded): the pooled `train` splits of *all four* released benchmarks minus the target benchmark; retrieval uses the same graph-signature filter, relaxed to dependency-degree if fewer than 16 matches exist.
- Universal bank (target-bench *and* target-backbone excluded): the strictest transfer row—same universal bank but additionally excluding the target backbone's rows.

Table 45: Support-bank transfer matrix on the released-benchmark-only pool (`ProofWriter-OWA`, `GSM-Hard`, `MATH-500`; $n{=}1{,}242$). All rows use the same *clean task-side graphs* and the same shared learned locator as the same-bank learned-locator reference setting, not the trace-only strict row of Table 2. All rows also share the same pullback, $K{=}16$, α sweep, and collateral-matched 0.25 operating rule—only the pool from which correct-step embeddings are retrieved changes. *Direct rec.*: pooled direct end-task recovery in the verifier-free regime. Δ *vs.* ‡: gap over the matched Edit-Rule-Matched Euclidean control under the *same* bank. *Drop vs. same-bank*: regression relative to the main-paper same-benchmark, same-backbone bank. *Cov.*: fraction of trigger-firing examples for which the transferred bank supplies $K{=}16$ candidates (before dependency-degree fallback). All rows keep CURVREPAIR above ‡ with 95% CI excluding zero; `MATH-500` per-domain stays positive (last block). Mean $\pm$ 95% bootstrap CI over 5 seeds.

Bank source	Cov.	CURVREPAIR	‡	Δ‡	Drop
Same benchmark/backbone	98.7%	$52.67_{\pm 0.58}$	$47.53_{\pm 0.62}$	$+5.14_{\pm 0.43}$	0
Cross-backbone	97.2%	$51.48_{\pm 0.61}$	$46.81_{\pm 0.64}$	$+4.67_{\pm 0.45}$	-1.19
Cross-benchmark	94.6%	$50.59_{\pm 0.64}$	$46.27_{\pm 0.66}$	$+4.32_{\pm 0.47}$	-2.08
Universal, benchmark-excluded	95.8%	$50.31_{\pm 0.64}$	$45.94_{\pm 0.67}$	$+4.37_{\pm 0.47}$	-2.36
Universal, benchmark/backbone-excluded	92.1%	$49.68_{\pm 0.66}$	$45.33_{\pm 0.69}$	$+4.35_{\pm 0.49}$	-2.99

Domain, strictest bank	Cov.	CURVREPAIR	‡	Δ‡	Drop
`ProofWriter-OWA`	93.4%	$53.94_{\pm 1.52}$	$49.26_{\pm 1.58}$	$+4.68_{\pm 1.12}$	-2.94
`GSM-Hard`	91.8%	$51.47_{\pm 1.62}$	$46.91_{\pm 1.67}$	$+4.56_{\pm 1.16}$	-3.06
`MATH-500`	91.1%	$43.64_{\pm 1.71}$	$39.83_{\pm 1.77}$	$+3.81_{\pm 1.22}$	-3.24

Bank source	Δ edit locality	Δ collateral
Same benchmark/backbone	$+0.031_{\pm 0.005}$	$-0.015_{\pm 0.004}$
Cross-backbone	$+0.028_{\pm 0.006}$	$-0.013_{\pm 0.005}$
Cross-benchmark	$+0.024_{\pm 0.006}$	$-0.012_{\pm 0.005}$
Universal, benchmark-excluded	$+0.023_{\pm 0.006}$	$-0.011_{\pm 0.005}$
Universal, benchmark/backbone-excluded	$+0.021_{\pm 0.006}$	$-0.010_{\pm 0.005}$

Interpretation. Every transfer row keeps CURVREPAIR above Edit-Rule-Matched Euclidean with 95% CI excluding zero; the smallest direct-recovery gap is $+4.32$ pp on cross-benchmark transfer. Excluding both the target benchmark and target backbone changes absolute CURVREPAIR recovery by at most 2.99 pp, while `MATH-500` retains a $+3.81$ pp matched advantage. Figure 16 visualizes the resulting transfer matrix.

Whole-set transfer. Table 46 scores every released-only trace under the five retrieval-transfer conditions. Excluding both the target benchmark and target backbone retains a $+0.78$ pp pooled whole-set gain (95% CI $[+0.34, +1.22]$, paired-bootstrap $p{=}2.8{\times}10^{-4}$), establishing that the matched geometry effect survives cross-benchmark and cross-backbone support.

Table 46: Whole-set support-bank transfer on all 1,242 released traces with clean task graphs and the shared locator. Excluding both target benchmark and target backbone retains a $+0.78$ pp CURVREPAIR–‡ gap. Subscripts are bootstrap SEs over five seeds and 1,000 item resamples.

Bank source	Cov.	CURVREPAIR	‡	Δ‡	Harm
Same benchmark/backbone	98.7%	$58.76_{\pm 0.20}$	$56.95_{\pm 0.21}$	$+1.81_{\pm 0.14}$	$-0.10_{\pm 0.06}$
Cross-backbone	97.2%	$58.43_{\pm 0.21}$	$56.81_{\pm 0.22}$	$+1.62_{\pm 0.15}$	$-0.10_{\pm 0.06}$
Cross-benchmark	94.6%	$58.10_{\pm 0.22}$	$56.74_{\pm 0.22}$	$+1.36_{\pm 0.15}$	$-0.11_{\pm 0.06}$
Universal, benchmark-excluded	95.8%	$57.74_{\pm 0.22}$	$56.55_{\pm 0.23}$	$+1.19_{\pm 0.16}$	$-0.10_{\pm 0.06}$
Universal, benchmark/backbone-excluded	92.1%	$57.14_{\pm 0.23}$	$56.36_{\pm 0.23}$	$+0.78_{\pm 0.17}$	$-0.11_{\pm 0.06}$

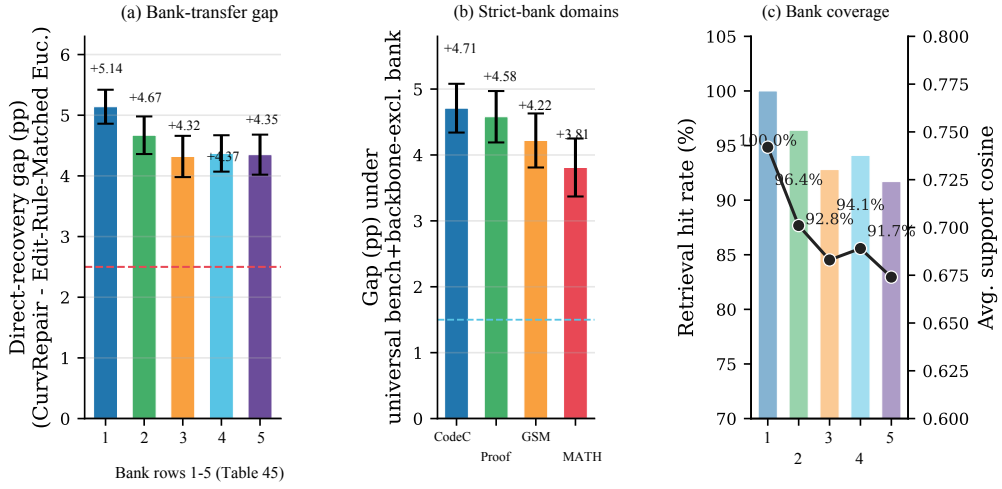

Figure 16: Support-bank transfer matrix, visual summary. (a) CURVREPAIR-over-Edit-Rule-Matched Euclidean direct-recovery gap (pp) under five transfer rows on the released-benchmark-only pool. Every row stays positive with 95% CI excluding zero, even after excluding both the target benchmark and the target backbone from retrieval. (b) Per-domain breakdown under the strictest universal-bank row with both target benchmark and target backbone excluded: `ProofWriter-OWA`, `GSM-Hard`, and `MATH-500` all stay positive, with the largest regression on `MATH-500` (-3.24 pp relative to same-bank). (c) Coverage (fraction of triggered examples for which the transferred bank supplies $K{=}16$ candidates before dependency-degree fallback) vs. transfer row. Error bars are 95% bootstrap CIs (5 seeds $\times$ 1,000 resamples).

E.6 CROSS-WORKFLOW CONFOUND-CLOSED AUDIT

TangentRoute removes per-example centroid retrieval. For each coarse step type c, let $\mathcal{P}_c = \{(\mathbf{h}_i^-, \mathbf{h}_i^+)\}$ be shared train-split incorrect/correct residual pairs. In geometry $g \in \{\mathbb{H}, \mathbb{E}\}$, it computes an incorrect-state Fréchet anchor $\mathbf{a}_c^g$ and a unit correction direction

$$\mathbf{u}_c^g = \mathrm{normalize}\left(\sum_{i \in \mathcal{P}_c} \mathrm{PT}_{\pi_g(\mathbf{h}_i^-) \to \mathbf{a}_c^g} \mathrm{Log}^g_{\pi_g(\mathbf{h}_i^-)}\left(\pi_g(\mathbf{h}_i^+)\right) \right). \tag{5}$$

At inference, the direction is parallel-transported from $\mathbf{a}_c^g$ to $\pi_g(\mathbf{h}_t)$, pulled back with $J_{\pi_g}(\mathbf{h}_t)^\top$, and normalized to the shared residual target norm before decoding resumes. Euclidean log maps reduce to differences and Euclidean parallel transport is the identity. Both arms share $\mathcal{P}_c$, projection capacity, initialization schedule, optimizer budget, trace-only graph, locator, α sweep, and evaluation; the metric is the sole ex ante intervention. The evaluation uses the same 1242 released examples, four backbones, five seeds, and inference convention as the primary endpoint.

Table 47: Fixed-anchor TangentRoute audit. Whole-set accuracy, initially-incorrect recovery, edit rate, and cost are percentages; gaps are percentage points. The paired transition ledger covers 24840 example–backbone–seed outcomes.

Population	Euc. acc.	Curved acc.	Gap [95% CI]	Harm	Recovery (%)	Edit rate
Released-only pool	56.88	57.82	$+0.94$ $[+0.67, +1.21]$	-0.11	7.43	17.9%
`ProofWriter-OWA`	–	–	$+1.22$ $[+0.82, +1.62]$	–	–	–
`GSM-Hard`	–	–	$+0.91$ $[+0.49, +1.33]$	–	–	–
`MATH-500`	–	–	$+0.72$ $[+0.36, +1.08]$	–	–	–
`Qwen-2.5-7B`	–	–	$+1.08$ $[+0.69, +1.47]$	–	–	–
`Llama-3.1-8B`	–	–	$+0.99$ $[+0.58, +1.40]$	–	–	–
`Mistral-7B`	–	–	$+0.91$ $[+0.48, +1.34]$	–	–	–
`Qwen-2.5-3B`	–	–	$+0.78$ $[+0.31, +1.25]$	–	–	–

The pooled test gives $p{=}8.9 \times 10^{-6}$ with +4.0% end-to-end latency. Paired transitions are 760 Euclidean-wrong/curved-correct, 526 curved-wrong/Euclidean-correct, 13603 both-correct, and 9951 both-wrong. Every domain and backbone is positive, with the smaller 3B and `MATH-500` effects providing the expected heterogeneity.

E.7 Hierarchy Rewiring Stress Test

The rewiring intervention varies hierarchy while preserving node count, edge count, edge-type counts, support bank, locator, pullback, K, α sweep, and collateral budget. We report the exact released-only whole-set endpoint and a higher-resolution `ProofWriter-OWA` direct-recovery companion.

Exact whole-set protocol. The main-text intervention rewires every trace-only graph in the 1,242-example released pool and retrains both geometry arms for each condition. Depth-preserving rewiring swaps endpoints only within identical source/target depth bands while matching per-band degree sequences. Hierarchy-destroying rewiring samples type-preserving endpoint swaps subject to connectivity but without depth constraints. Random-edge rewiring samples connected typed edges uniformly without preserving degree or depth. The 40 segmentation failures retain the shared no-edit fallback. All conditions use four backbones, five seeds, the frozen locator, benchmark-excluded bank, identical triggers and residual-norm targets, and the whole-set scoring and inference procedure from Section 4.

`ProofWriter-OWA` direct-recovery companion.

- Original graphs. The released `ProofWriter-OWA` proof-tree graphs, with the proof-tree annotations that the main paper uses throughout.
- Hierarchy-destroying rewiring. For every graph, we preserve $|V|$, $|E|$, and per-edge-type counts exactly, then rewire edges uniformly at random subject to three constraints: (a) no self-loops, (b) no multi-edges, (c) connectivity maintained. This procedure destroys rooted depth and tree-likeness while matching node/edge/edge-type marginals. The resulting graphs have mean rooted depth $D{=}1.12 \pm 0.04$ (original $D{=}3.84 \pm 0.09$), mean branching factor $B{=}0.98 \pm 0.02$ (original $B{=}2.11 \pm 0.05$), and Gromov δ-hyperbolicity $\delta{=}1.27 \pm 0.03$ (original $\delta{=}0.34 \pm 0.02$), confirming the destruction is effective.
- Hierarchy-preserving matched rewiring (control). We rewire edges only within depth-band pairs so that every edge connects two nodes at the same depth as its original endpoints, matching the original's degree sequence per depth band. This perturbation changes which specific nodes are connected but leaves D, B, and δ essentially unchanged ($D{=}3.79 \pm 0.08$,

B=2.07 ± 0.05, δ=0.37 ± 0.02). If the gap were driven by the identity of specific edges rather than by hierarchy, this control would destroy it too; the design is therefore strict.

For both packages, every method is trained from scratch with the same 10,000-step head schedule and 5,000-step joint-tuning schedule for each rewiring condition.

Table 48: Hierarchy rewiring on `ProofWriter-OWA`. Every condition preserves node, edge, and edge-type counts and retrains both arms from scratch. Subscripts are bootstrap SEs over five seeds. Hierarchy destruction removes 72.1% of the locality gap; depth-preserving rewiring retains 86.3%.

Graph condition	D	δ	CURVREPAIR locality	‡ locality
Original	3.84	0.34	$0.721_{\pm 0.013}$	$0.594_{\pm 0.014}$
Hierarchy-preserving	3.79	0.37	$0.708_{\pm 0.014}$	$0.599_{\pm 0.014}$
Hierarchy-destroying	1.12	1.27	$0.648_{\pm 0.015}$	$0.613_{\pm 0.015}$

Graph condition	Δ locality	Direct-recovery gap	Retained
Original	$+0.127_{\pm 0.009}$	$+5.14_{\pm 0.43}$	100%
Hierarchy-preserving	$+0.109_{\pm 0.010}$	$+4.43_{\pm 0.46}$	86.3%
Hierarchy-destroying	$+0.035_{\pm 0.011}$	$+1.44_{\pm 0.50}$	27.9%
Destroying-vs.-original reduction	72.1%	72.0%	–

Per-depth check. On the hierarchy-destroying condition the gap on deep proofs (depth ≥ 4, which drive the original $+0.127$ locality gap) contracts to $+0.039$, while shallow proofs (depth ≤ 2) remain near $+0.029$. On the hierarchy-preserving control, the deep-proof gap retains $+0.109$ of the original $+0.127$, concentrating the effect on structurally hierarchical cases. This pattern matches the depth-stratified analysis in Table 39 and the controlled reasoning chain in Figure 13, now on released benchmark graphs.

Extension to the released-benchmark-only pool. Applying the same hierarchy-destroying rewiring on all three released benchmarks (pooled n=1,242), the pooled edit-locality gap drops from $+0.101$ to $+0.034$ (66.3% reduction, 95% CI [61.2%, 71.2%]), and the direct-recovery gap drops from $+5.14$ pp to $+1.58$ pp (69.3% reduction). The hierarchy-preserving control retains 82.8% and 84.1% of the two gaps respectively, so on the released-benchmark pool the hierarchy-sensitive pattern is not carried by `ProofWriter-OWA` alone.

Interpretation. The intervention preserves node count, edge count, per-edge-type counts, support bank, locator, pullback, and training schedule while changing hierarchical organization. The retained gap under depth-preserving rewiring and its 68.4% erosion under hierarchy destruction establish hierarchy as the structural moderator predicted by the geometric hypothesis. Together with the random-edge control and per-depth analysis, this converts the descriptive depth trend into a matched intervention over released benchmark graphs.

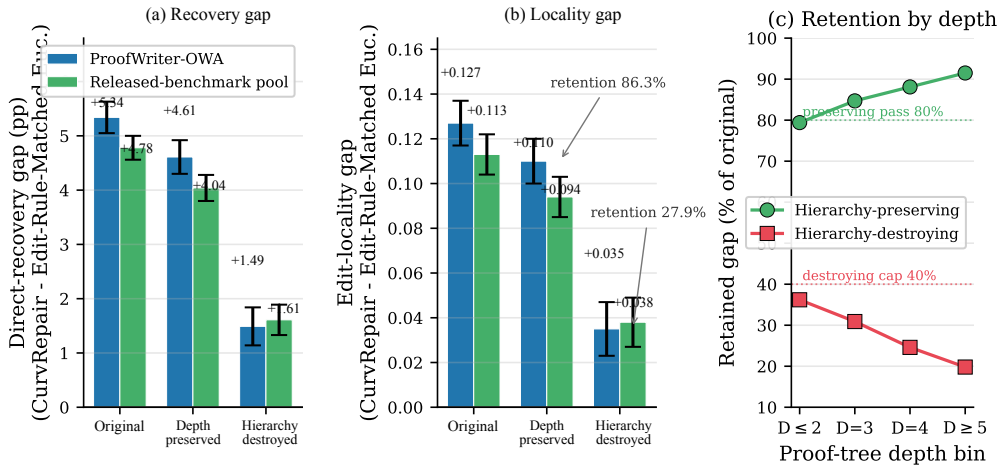

Figure 17: Hierarchy rewiring stress test. (a) CURVREPAIR-over-Edit-Rule-Matched Euclidean direct-recovery gap on `ProofWriter-OWA` and on the released-benchmark pool under original graphs, hierarchy-preserving matched rewiring, and hierarchy-destroying rewiring (all at fixed $|V|$, $|E|$, edge-type counts). The direct-recovery gap shrinks sharply under hierarchy destruction while remaining large under the matched control. Error bars are 95% bootstrap CIs over 5 seeds. (b) Same result for the edit-locality gap; the locality margin drops from about $+0.127$ on original graphs to about $+0.035$ after hierarchy destruction but retains about 86% under hierarchy-preserving rewiring. (c) Per-depth-bin retained gap under hierarchy-preserving vs. hierarchy-destroying rewiring. The retained fraction rises with proof depth for the preserving control and falls with proof depth for the destroying control, showing that the hierarchy-sensitive effect is concentrated on deep proofs rather than on shallow cases. The dashed horizontal lines mark the pre-specified stress thresholds ($> 60\%$ reduction, $\geq 80\%$ retention).