# DO LANGUAGE MODELS REVISE THE RIGHT CONCLUSIONS? DIAGNOSING AND TRAINING EVIDENCE-CONTINGENT REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Aggregate answer scores cannot tell whether a language model revises the right conclusions when its evidence changes. We study *evidence-contingent reasoning* (ECR): withdrawing conclusions that lose support, preserving unaffected content, and retaining conclusions that have a valid alternative proof in the target world. Our evaluation suite, EviPair, links audited evidence worlds and measures Change, Preserve, and Switch on fixed, model-independent eligibility sets. On MuSiQue-Full, FEVER, and QAMPARI, reinforcement learning with only node-local semantic and evidence rewards improves answering but lowers preservation on every task, and development-set endpoint matching still leaves substantial response differences on the full test graphs. We instantiate structured intervention training with Paired Transition Policy Optimization (PTPO), which credits each node with the marginal return of its incident intervention edges while reusing independently sampled world trajectories. With Forced Switch edges (which invalidate every source-world proof of a conclusion while retaining alternative support) withheld from reinforcement learning, macro SwitchAcc on FEVER and QAMPARI rises from 17.48 after supervised finetuning (SFT) to 33.42. The gain over the strongest prior-method baseline, CRPO-Evidence, is 4.28 points (95% paired root-clustered interval [2.21, 6.35]); the Change and Preserve differences are not significant. The clearest benefit of structured intervention training over CRPO-Evidence thus lies in alternative-proof adaptation, and local revision and proof adaptation are best evaluated as separate targets.

## 1 INTRODUCTION

A correct answer does not specify what should happen when its evidence changes. Consider a question with three supported conclusions. Removing the only proof of one conclusion should withdraw that conclusion, not erase the other two. Removing an irrelevant passage should leave the answer intact. Replacing the original proof with a different valid proof should preserve the conclusion but change its justification. These obligations concern *which* content changes and *which* evidence sustains it, not merely whether two answer strings differ (Figure 1).

Behavioral tests distinguish directional change from invariance (Ribeiro et al., 2020), and evidence interventions expose disconnected reasoning (Lee et al., 2021; Guo et al., 2023; Yun et al., 2026). CRPO further trains these behaviors through cross-branch rewards (Du et al., 2026). These foundations motivate a more specific question: can a model withdraw an unsupported conclusion without losing supported content, and retain a conclusion when its justification must change?

We study this capability as *evidence-contingent reasoning* (ECR). Our evaluation suite, EviPair, links independently evaluated evidence worlds through audited interventions and reports Change, Preserve, and Switch separately. The evaluation object is the aligned pair of outputs and their support obligations, rather than an aggregate endpoint score or an answer-string difference. Eligibility is fixed before evaluating any model. Forced Switch, for example, invalidates every complete source-world proof of a designated target while retaining alternative support; source errors and missing citations remain failures.



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

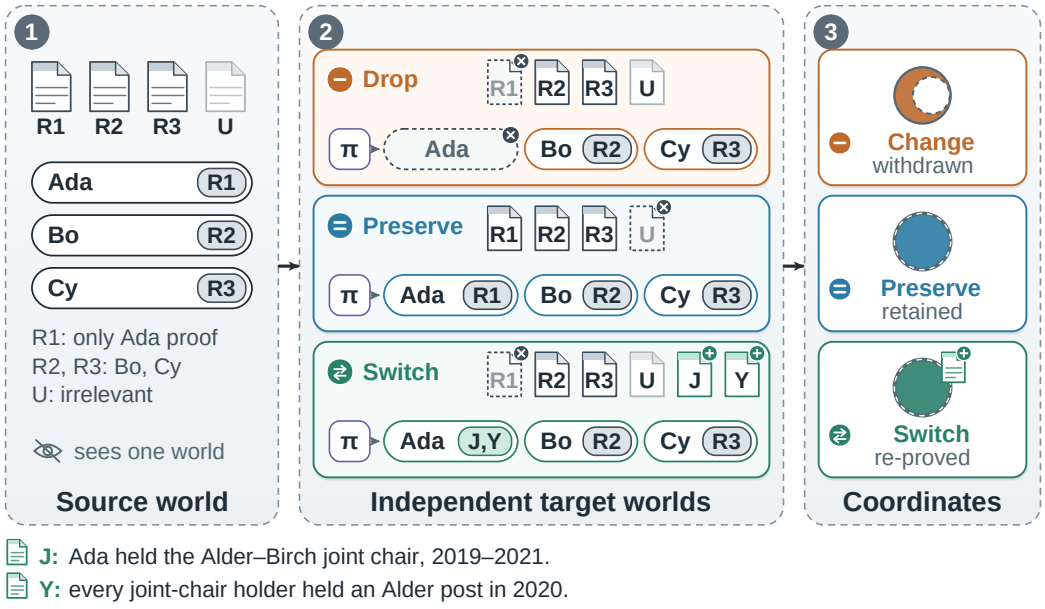


Figure 1: **Change exactly what the evidence warrants.** A fictional multi-answer question ("Who held an Alder Institute post in 2020?") with valid outputs, not model predictions. (1) The source world holds three answer-supporting documents and one irrelevant one; each answer cites its proof. (2) Three independent target worlds, each answered by the same policy $\pi$ without access to the other worlds: Drop removes Ada's only proof; Preserve removes the irrelevant document; Switch replaces R1 by the joint proof {J,Y} (dashed with $\times$: removed; green with $+$: added). (3) Each world probes one coordinate: solid circle = source answers, dashed circle = target answers, fill = the part counted (withdrawn for Drop→Change; retained for Preserve; retained with a new proof for Switch). For Preserve and Switch the two sets coincide.

The same setting defines the training question. We use Paired Transition Policy Optimization (PTPO) to train on typed interventions while withholding Forced Switch edges during reinforcement learning (RL). PTPO is an edge-additive implementation with marginal edge credit, not a higher-order graph objective. This setting tests whether training on withdrawal, preservation, and natural redundancy improves proof adaptation to the held-out relation.

Our contributions map onto the sections that follow. **Evaluation** (Section 3): EviPair separates withdrawal, retention, and proof adaptation on gold-defined, model-independent eligibility sets, including failures in either world. **Diagnosis** (Section 4): endpoint-only reinforcement (GRPO-Marginal) improves answering over its supervised-finetuned (SFT) start but lowers preservation on all three tasks (Figure 2), and development-matched endpoints still leave substantial response differences. **Training** (Sections 5–6): PTPO, an edge-additive estimator that credits each node with the marginal return of its incoming and outgoing intervention edges, is tested against this diagnosis under the held-out Forced Switch relation.

## 2 Related Work

**Behavior, evidence, and counterfactual evaluation.** CheckList motivates testing capabilities beyond aggregate accuracy (Ribeiro et al., 2020). Evidence perturbations and counterfactual multihop inference expose shortcuts (Lee et al., 2021; Guo et al., 2023; Yun et al., 2026). MuSiQue, FEVER, and QAMPARI provide complementary multihop, claim-verification, and multi-answer settings (Trivedi et al., 2022; Thorne et al., 2018; Amouyal et al., 2023); their original annotations do not by themselves establish exhaustive alternative proofs in our bounded worlds. Citation evaluation also distinguishes answer quality from supporting evidence (Gao et al., 2023). ECR combines local set-valued obligations with a model-independent proof-switching denominator. Valid citations are observable evidence use, not proof of faithful latent reasoning.

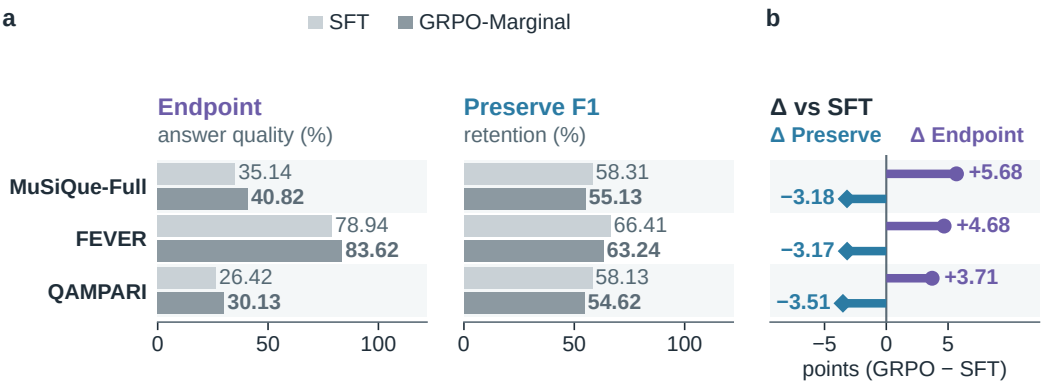


Figure 2: **Better answering can coincide with worse preservation.** (a) Endpoint (violet axis) and Preserve F1 (blue axis) of the SFT start (light gray) and GRPO-Marginal (dark gray) on each task; five-seed means on each metric's own 0–100 scale (SDs are at most 0.6 and omitted). (b) Differences of these means (GRPO-Marginal minus SFT): Endpoint rises and Preserve falls on every task. Marks are differences of reported means, not uncertainty intervals; endpoint metrics are task-native, so no cross-task endpoint aggregate is formed. Complete values remain in Table S27.

**Relational rewards and policy credit.** PTPO uses the clipped surrogate and group-based credit conventions of PPO, GRPO, and REINFORCE-style optimization (Schulman et al., 2017; Shao et al., 2024; Ahmadian et al., 2024). CRPO averages relational outcomes over a neighboring branch (Du et al., 2026); its CRPO-Evidence variant uses a binary semantic relation kernel with joint standardization, whereas PTPO uses typed relation returns, credit from incoming and outgoing edges, and node-local leave-one-out baselines, each of which we isolate (Table S26). Search-R1, Knowledgeable-R1, and DPEPO address search-enabled reinforcement, contextual interference, and parallel exploration, respectively (Jin et al., 2025; Lin et al., 2026; Zhang et al., 2026). COMA offers a related credit-assignment perspective but marginalizes an agent's own actions rather than a neighboring world's trajectories (Foerster et al., 2017). We therefore compare reward specification, credit scaling, pairing, and rollout reuse separately, instead of attributing their combined effect to graph structure.

## 3 Evidence-Contingent Reasoning

We now formalize the worlds, interventions, and coordinates used throughout.

**Worlds and proofs.** A root problem $r$ specifies a question $q$ and an audited finite evidence universe. A world $w_v$ contains available evidence $E_v$ and a tool-visibility configuration. Independently sampling the policy in each world produces a trajectory $\tau_v$ and parsed output $o_v = (Y_v, C_v)$: a set of semantic atoms and a certificate for each predicted atom. Let $\mathcal{S}_r(a)$ be the audited family of sufficient evidence sets for conclusion $a$. Then

$$\text{Supported}(a, w_v) \iff \exists S \in \mathcal{S}_r(a) : S \subseteq E_v. \tag{1}$$

Each proof may require several facts jointly. Redundant supersets are removed before determining eligibility. A valid certificate asserts a correct atom, cites only available and observable sources, and contains a complete proof. Atoms follow an audited task-specific mapping: supported sub-conclusions for MuSiQue, claim–label units for FEVER, and answer entities for QAMPARI (Appendix A.1). Support is relative to the declared finite universe, not just the original dataset annotations.

**Typed interventions.** An Evidence Intervention Graph $\mathcal{G}_r = (V_r, \mathcal{A}_r)$ has directed edges $e = (u, v, \rho_e, \Gamma_e)$, where $\rho_e$ is the primary relation and $\Gamma_e$ contains audited atomic obligations. Drop removes support for specified conclusions; Preserve leaves relevant support intact; Switch preserves semantics while invalidating all old complete proofs for specified targets. Selective set removal can require Drop and Preserve on the same edge. The constraints are derived from gold semantics and proof structure, not from a possibly incorrect source prediction. They define a relation $\mathcal{R}_e(o_u, o_v; \Gamma_e)$,

not a unique successor string. All graph metadata, gold labels, and the other world's outputs are hidden from the policy. Here, "revision" denotes a relation between independently generated outputs, not sequential editing or stateful self-correction.

**Three evaluation coordinates.** For gold sets $A_u, A_v$ and predictions $Y_u, Y_v$, define

$$D_e^* = A_u \setminus A_v, \quad P_e^* = A_u \cap A_v, \qquad \widehat{D}_e = Y_u \setminus Y_v, \quad \widehat{P}_e = Y_u \cap Y_v. \tag{2}$$

**Change** measures withdrawal: the F1 of $\widehat{D}_e$ against nonempty $D_e^*$. **Preserve** measures retained content: the F1 of $\widehat{P}_e$ against nonempty $P_e^*$. These eligibility sets can overlap; an edge may require withdrawing one atom while retaining another. Constant empty answers fail preservation, and constant wrong answers incur false positives. Parse failure is a separate state $\perp$, scored zero rather than interpreted as an empty answer. Table S21 separates precision and recall.

**SwitchAcc** uses a pre-frozen set of Forced Switch edges. The gold answer set is unchanged; for each designated target no complete source proof survives, but a target-world proof exists. An edge succeeds only if all designated targets are correctly retained with valid certificates in both worlds, no invalid source is cited, and neither output contains extra incorrect semantic predictions. Missing other correct answers is still penalized by Preserve and Endpoint. Every eligible edge stays in the denominator, including source errors, absent citations, and tool failures. Forced Switch generally requires evidence replacement or addition; it is not pure deletion.

Natural redundancy has a different obligation. If deleting one proof leaves another intact, a model already using the surviving proof should not be forced to switch. We therefore report fixed-target SupportAdaptAcc and a separate triggered-unit diagnostic with coverage (Table S9). Unexpected additions, newly unsupported answers, and endpoint precision accompany the main coordinates (Table S6); T-Exact is auxiliary strict joint reliability (Table S5). Although ECR is computed from complete endpoint outputs, aggregate endpoint means do not identify the instance-aligned response (Appendix B).

## 4 EviPair and Behavioral Diagnostics

**Endpoint improvement can accompany worse preservation.** The endpoint-only baseline, GRPO-Marginal, receives node-local semantic and certificate rewards but no cross-world relation reward. It gains 5.68, 4.68, and 3.71 endpoint points over SFT on MuSiQue, FEVER, and QAMPARI, respectively, while losing 3.18, 3.17, and 3.51 preservation points (Table S27). Its macro changes are $+7.12$ Change, $-3.28$ Preserve, and $+2.56$ Switch, with SwitchAcc reaching 20.04. The diagnostic is the preservation loss despite improved answering, not the gap between endpoint and switching metrics. Base's 14.91% parse-failure rate is reported separately rather than treated as a reasoning deficit.

Error decomposition is consistent with excessive revision: GRPO-Marginal has FalseDrop 14.83 versus SFT's 9.72, and NewUnsupported 13.21 versus 8.92 (Table S6). Conditional on a correct source answer, over-retraction is 24.91 for GRPO-Marginal and 13.41 for PTPO (Table S12); those method-dependent subsets are diagnostic only, and the main comparisons retain the frozen full denominators.

**Setting and evaluation units.** EviPair derives audited worlds from MuSiQue-Full ("MuSiQue" below), FEVER, and bounded-shard QAMPARI. MuSiQue and FEVER use static evidence; QAMPARI uses bounded evidence with direct-context and search/open/final evaluations reported separately. Roots are split before world construction. Training contains 12,136 roots, 41,098 worlds, and 36,243 typed edges; test contains 1,548 roots, 5,241 worlds, and 4,622 edges (Table S2). The frozen denominators are 1,971 Change edges, 3,185 Preserve edges, and 485 Forced Switch edges (Table S4). MuSiQue has no audited Forced Switch family and is excluded, not zero-filled, in the Switch macro.

Endpoint retains task-specific meaning: sufficient-node answer F1 for MuSiQue, atomic label recall on derived FEVER worlds, and shard-bounded answer-set F1 for QAMPARI. Nodes are deduplicated and averaged within roots before root averaging. Change/Preserve average edges within root and primary type, then applicable roots and pre-defined types equally. SwitchAcc retains its literal fixed-edge success ratio. Cross-task macros use all three tasks for Change/Preserve and only

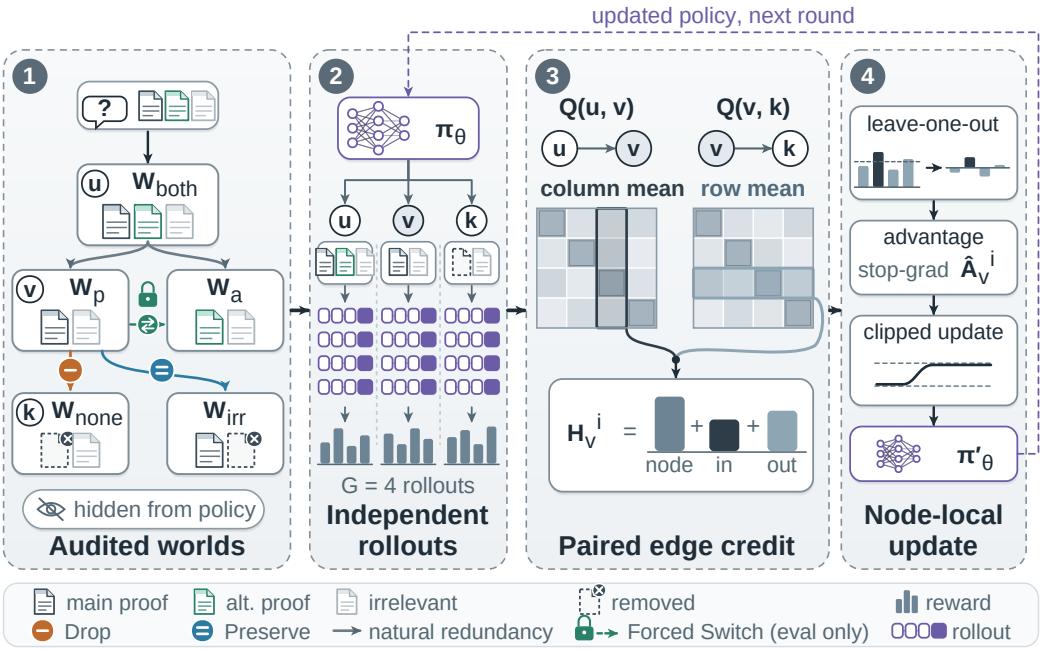


Figure 3: **PTPO overview.** (1) Audited evidence-world graph per root. Documents: main proof (dark), alternative proof (green), irrelevant (light); dashed with $\times$: removed. Edges: Drop (orange), Preserve (blue), natural redundancy (gray); the dashed green lock edge is Forced Switch, evaluation-only and withheld from RL. Types and labels are hidden from the policy; $u, v, k$ mark the path used in (3). (2) The policy $\pi_\theta$ sees only its own world; each row is one of $G = 4$ rollouts, followed by its node reward (bars). (3) For $e = (u, v)$, returns $Q^e_{ij}$ (all $G \times G$ pairs incl. $i = j$, outlined) are averaged over columns (incoming) and rows (outgoing) and added to the node reward: the edge-additive credit $H^i_v$. (4) Node-local leave-one-out yields the advantage $\hat{A}^i_v$ (stop-gradient) for the clipped update; the updated policy is used next round.

FEVER/QAMPARI for Switch; Endpoint is not used as a mixed-scale main ranking. Full definitions and companion diagnostics are in Appendix A.

**Uncertainty and inference.** Main-table $\pm$ values describe five-seed sample standard deviations, not inferential uncertainty. Inference uses paired root-clustered bootstrap intervals with 10,000 resamples per seed, retaining all worlds and edges of each resampled root (Dror et al., 2018). Switch is the sole pre-registered primary family, with six Holm–Bonferroni-corrected comparisons; other tests are exploratory (Tables S3, S16, and S17).

## 5 PTPO: Structured Intervention Training

**A normalized, edge-additive objective.** Section 4's diagnosis motivates crediting each node by its intervention outcomes, not only its endpoint. Figure 3 summarizes the pipeline. Let $a_v = 1/|V_r|$ and $b_e = 1/(|T_r| n_{\rho_e})$, where $T_r$ contains the applicable primary types and $n_{\rho_e}$ counts edges of the type of $e$. Both sets of weights sum to one within a root, and $\lambda \in [0, 1]$ weights the edge term against the node term. We optimize

$$J(\theta) = \mathbb{E}_r \left[ (1 - \lambda) \sum_v a_v \mathbb{E} R_v(\tau_v) + \lambda \sum_{e=(u,v)} b_e \mathbb{E} R_e(\tau_u, \tau_v; \Gamma_e) \right]. \tag{3}$$

The node reward balances semantic correctness and valid evidence and is counted once per node, not per incident edge. The edge reward averages applicable withdrawal, preservation, and support-adaptation returns, gated by format validity and both nodes' semantic precision (Appendix A.6). Applicability is gold-defined. In the reported RL curriculum, natural redundancy supplies alternative-support supervision; Forced Switch edges and their corresponding reward component are absent.

**Marginal edge credit.** Sample $G \geq 2$ independent trajectories at each node and let $Q^e_{ij} = R_e(\tau^i_u, \tau^j_v)$ contain only the relation return. For node $v$, PTPO computes

$$H^i_v = (1-\lambda) a_v R_v(\tau^i_v) + \lambda \sum_{e=(v,k)} \frac{b_e}{G} \sum_j Q^e_{ij} + \lambda \sum_{e=(k,v)} \frac{b_e}{G} \sum_j Q^e_{ji}, \tag{4}$$

$$\hat{A}^i_v = \operatorname{stopgrad}\left( H^i_v - \frac{1}{G-1} \sum_{k \neq i} H^k_v \right). \tag{5}$$

Both incoming and outgoing edges contribute, while each node's trajectory group is generated once and reused across incident edges. The node-local leave-one-out (LOO) baseline centers these credits before the clipped token-ratio update. Under independent node sampling and no self-loops, the credit estimator is unbiased at the rollout policy (Appendix B). This statement does not extend to active clipping or repeated off-policy updates. The full loss and token-masking conventions are given in Eq. (35) of Appendix B.

**What full pairing does and does not add.** PTPO-Shared replaces full pairing with random one-to-one matching using the same node samples and LOO convention. Full pairing evaluates $|\mathcal{A}_r|G^2$ correlated combinations from $|V_r|G$ generated trajectories; it does not create extra independent observations. Conditional averaging gives non-increasing covariance for the raw matching estimator, before baseline and clipping transformations (Appendix B). A degree-corrected decomposition that samples each typed pair independently (Typed-Pair) also estimates the same objective. Our controls therefore test the effects of reward, credit, pairing, and rollout reuse separately.

## 6 Experiments

**Models and controls.** The main backbone is Qwen3-4B-Base, with Llama-3.1-8B-Instruct as a confirmatory family. Within each of five seeds $\{17, 42, 73, 101, 2024\}$, all RL methods start from the same Stage-A checkpoint (single-world supervised finetuning in the interface format). RL uses $G = 4$, $\lambda = 0.50$, clipping $\epsilon = 0.20$, learning rate $10^{-6}$, and 2,000 steps on four A40 GPUs; evaluation is greedy. Endpoint-LOO isolates endpoint reward under PTPO's credit convention. CRPO-Evidence sees the same training edges with a binary semantic relation kernel and its stated joint-standardization adaptation. Typed-Pair and PTPO-Shared test independent edge sampling and random matching. Frozen hyperparameters, budgets, and implementation checks appear in Tables S1, S8, and S13.

### 6.1 Structured training improves held-out proof switching

Despite excluding Forced Switch edges from RL, PTPO reaches macro Change/Preserve/Switch of 57.75/66.58/33.42, compared with SFT's 44.63/60.95/17.48 (Table 1). Relative to CRPO-Evidence, the Switch gain is $+4.28$ with paired root-clustered 95% interval $[2.21, 6.35]$ and $p_{\mathrm{Holm}} = 1.02 \times 10^{-4}$. The corresponding Change and Preserve intervals include zero (Figure 4). Positive differences at each training seed do not override these root-level intervals.

Relative to CRPO-Evidence, the advantage is confined to Switch. Endpoint values remain close to those of GRPO-Marginal, but this is not a blanket non-inferiority result: the PTPO–GRPO-Marginal interval extends to $-1.12$, below the $-1.00$ margin (Table S3).

### 6.2 Behavior differs at similar endpoint levels

Checkpoints are selected independently per seed on development data, never on test response scores. MuSiQue simultaneously matches sufficient-node answer F1 and answerability on both sufficient and insufficient nodes; FEVER and QAMPARI match their semantic endpoint. The pre-registered matching tolerances $\delta_{\mathrm{match}}$, comparability threshold $\kappa = 1.50$, and non-inferiority margin $\eta = 1.00$ are frozen (Table S1). All selected checkpoints are evaluated on the complete frozen test graphs, and the selected steps with their development metrics are recorded in Table S23.

The full-test endpoint ranges are 0.60, 0.39, and 0.68 points for MuSiQue, FEVER, and QAMPARI, respectively (Table S28). Within these ranges, PTPO exceeds GRPO-Marginal by 4.78 Change

Table 1: **Training on frozen intervention graphs.** Five-seed mean ± SD, 0–100, higher is better. Endpoint is task-native; Switch (SwitchAcc) covers FEVER and QAMPARI. The Δ rows are differences of the macro means from unrounded per-seed values (paired intervals: Table S3). A dash marks an inapplicable entry (no MuSiQue Forced Switch family; no cross-task Endpoint aggregate). Bold and green mark the focal method, not the best or a significant result. QAMPARI uses static direct context (retrieval agents: Table S14).

| Method | Endpoint | Change | Preserve | Switch |
|---|---|---|---|---|
| *Macro (average over tasks; Switch: FEVER and QAMPARI)* | | | | |
| SFT | — | 44.63 ±0.25 | 60.95 ±0.24 | 17.48 ±0.38 |
| GRPO-Marginal | — | 51.75 ±0.19 | 57.67 ±0.20 | 20.04 ±0.33 |
| CRPO-Evidence | — | 56.96 ±0.23 | 65.92 ±0.18 | 29.15 ±0.49 |
| Endpoint-LOO | — | 52.34 ±0.33 | 60.29 ±0.29 | 20.65 ±0.48 |
| Typed-Pair | — | 55.81 ±0.57 | 64.62 ±0.42 | 27.94 ±0.74 |
| PTPO-Shared | — | 56.48 ±0.37 | 65.46 ±0.30 | 29.73 ±0.60 |
| **PTPO** | — | 57.75 ±0.26 | 66.58 ±0.24 | 33.42 ±0.38 |
| Δ PTPO − SFT | — | +13.12 | +5.62 | +15.94 |
| Δ PTPO − CRPO-Evidence | — | +0.79 | +0.65 | +4.28 |
| *MuSiQue-Full* | | | | |
| SFT | 35.14 ±0.23 | 44.72 ±0.26 | 58.31 ±0.26 | — |
| GRPO-Marginal | 40.82 ±0.33 | 52.34 ±0.39 | 55.13 ±0.42 | — |
| CRPO-Evidence | 40.13 ±0.31 | 56.42 ±0.54 | 62.81 ±0.44 | — |
| Endpoint-LOO | 40.41 ±0.27 | 52.87 ±0.42 | 57.64 ±0.32 | — |
| Typed-Pair | 40.58 ±0.35 | 55.42 ±0.53 | 61.62 ±0.39 | — |
| PTPO-Shared | 40.29 ±0.29 | 56.13 ±0.50 | 62.38 ±0.45 | — |
| **PTPO** | 40.71 ±0.31 | 57.24 ±0.39 | 63.52 ±0.41 | — |
| *FEVER* | | | | |
| SFT | 78.94 ±0.22 | 51.37 ±0.24 | 66.41 ±0.26 | 19.69 ±0.39 |
| GRPO-Marginal | 83.62 ±0.25 | 58.71 ±0.40 | 63.24 ±0.35 | 22.50 ±0.58 |
| CRPO-Evidence | 83.87 ±0.34 | 64.83 ±0.43 | 71.24 ±0.45 | 31.63 ±0.66 |
| Endpoint-LOO | 83.21 ±0.30 | 59.28 ±0.39 | 65.83 ±0.39 | 23.13 ±0.55 |
| Typed-Pair | 83.41 ±0.36 | 63.61 ±0.56 | 69.91 ±0.38 | 30.31 ±0.59 |
| PTPO-Shared | 83.13 ±0.34 | 64.21 ±0.39 | 70.87 ±0.33 | 32.19 ±0.56 |
| **PTPO** | 84.02 ±0.30 | 65.61 ±0.41 | 71.83 ±0.36 | 35.94 ±0.50 |
| *QAMPARI* | | | | |
| SFT | 26.42 ±0.31 | 37.81 ±0.38 | 58.13 ±0.33 | 15.27 ±0.49 |
| GRPO-Marginal | 30.13 ±0.44 | 44.21 ±0.58 | 54.62 ±0.60 | 17.58 ±0.90 |
| CRPO-Evidence | 29.62 ±0.54 | 49.63 ±0.77 | 63.72 ±0.56 | 26.67 ±0.91 |
| Endpoint-LOO | 29.81 ±0.38 | 44.87 ±0.64 | 57.41 ±0.56 | 18.18 ±0.75 |
| Typed-Pair | 30.04 ±0.53 | 48.41 ±0.73 | 62.34 ±0.57 | 25.58 ±0.94 |
| PTPO-Shared | 29.71 ±0.43 | 49.12 ±0.66 | 63.14 ±0.49 | 27.27 ±1.00 |
| **PTPO** | 30.28 ±0.41 | 50.41 ±0.71 | 64.38 ±0.63 | 30.91 ±0.82 |

and 7.72 Preserve points on MuSiQue, and by 10.69/13.34 Switch points on FEVER/QAMPARI. Figure 5 places the dev-selected checkpoints within the measured endpoint–response trajectories. Thus, measured aggregate endpoint quality does not summarize the response differences. Matching does not remove differences in optimization history or instance-level errors and is not a randomized causal intervention.

### 6.3 Transfer to held-out relations and structures

**Forced Switch is held out during RL.** The curriculum contains Necessary/Selective Drop, Irrelevant Preserve, and Natural Redundancy, but no Forced Switch edges. Holdouts operate at the root level: no world or edge from a held-out root enters the RL graph. The macro SwitchAcc gain reported

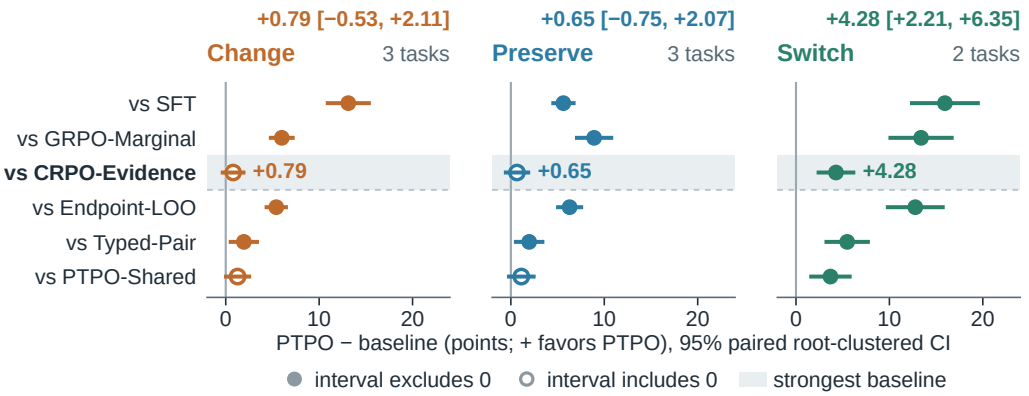


Figure 4: **The clearest comparative gain is proof adaptation.** PTPO minus each baseline on Change, Preserve, and Switch with paired root-clustered 95% intervals (Table S3). The strongest prior-method baseline, CRPO-Evidence, is shaded; the numbers above each panel give its difference and interval. Filled markers: the interval excludes zero; hollow: it includes zero. Change/Preserve are exploratory. Holm correction applies to $p$-values in the pre-registered Switch family, not to these intervals. Endpoint contrasts are in Table S3(b); Switch contrasts with $p$-values are in Table S29.

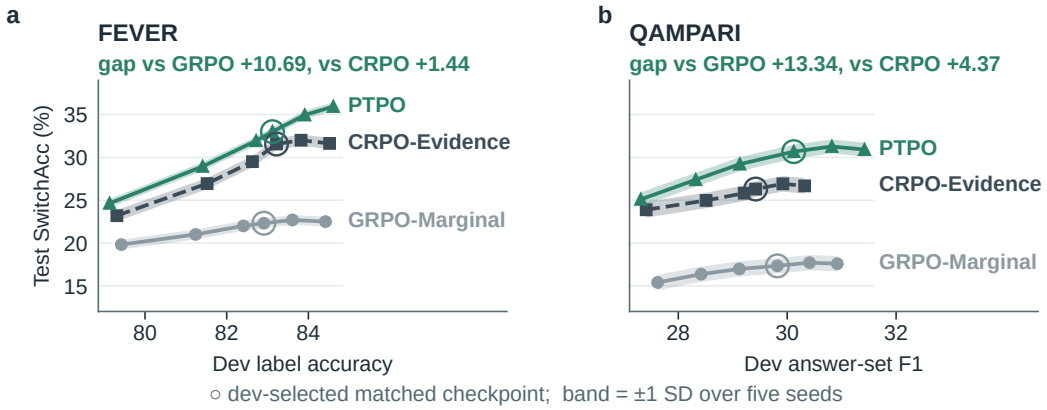


Figure 5: **Endpoint–switching response curves** on (a) FEVER and (b) QAMPARI. Each marker is an evaluated checkpoint: dev endpoint on the horizontal axis and full-test SwitchAcc on the vertical. Rings identify the dev-selected matched checkpoint of Table S28; bands show five-seed standard deviations, not inferential intervals. The printed gaps are PTPO minus GRPO-Marginal and minus CRPO-Evidence at those checkpoints; each curve's rightmost checkpoint is the Table 1 model. MuSiQue is omitted because Switch is inapplicable.

above (17.48 to 33.42) is consistent with transfer to this RL-held-out relation. Removing natural-redundancy edges together with their dedicated alternative-support nodes lowers switching to 28.99 (Table S25(a); Figure 6 shows the task-wise results). The 4.43-point difference concerns this joint removal, not the reward component alone. On these retained edges, PTPO's macro SupportAdaptAcc is 67.75 against SFT's 52.22 (Table S9(a)). Stage-A remains shared single-world supervision over the training roots; the holdout applies to RL, not all finetuning or pretraining exposure.

**Topology, task, and backbone checks.** The topology holdouts show directional gains but lower absolute performance: Change is 53.12 for joint multi-sentence proofs and 44.92 for long-hop questions, while Switch is 28.59 under the held-out per-target predicate of at least three alternative proofs (Table S25(b)). The cross-task direction is MuSiQue → QAMPARI only, under the unified static direct-context interface with target data excluded from finetuning and selection (Table S25(c)). Agent transfer is separate (Table S14); static–agent differences are not interpreted as a controlled effect. Llama confirmation reproduces the preservation pattern: GRPO-Marginal lowers MuSiQue

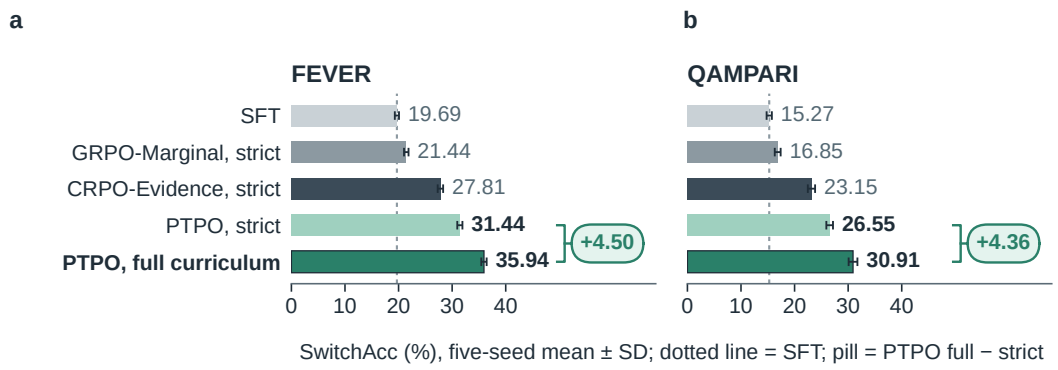


Figure 6: **Transfer without Forced Switch training edges** on (a) FEVER and (b) QAMPARI. Bars show five-seed means with ± SD whiskers, not confidence intervals; the dotted line marks the SFT start. Strict curricula remove natural-redundancy edges together with their dedicated alternative-support nodes. The bracketed value is PTPO full minus strict; only this comparison varies the curriculum within one method. Complete values, including the macro, are in Table S25(a).

Preserve from 60.12 to 57.21, whereas PTPO reaches 65.42 (Table S25(d)). These checks do not establish builder-independent generalization.

### 6.4 Reward, estimator, and reuse effects

Changing the binary kernel to typed returns under joint-standardized credit raises FEVER/QAMPARI Switch from 31.63/26.67 to 33.13/28.24, and changing the credit convention to LOO yields 35.94/30.91 (Table S26); these are separate interventions, so their combined gain is not attributed to either factor alone. Full pairing has higher observed Switch than random matching, but its Change/Preserve contrasts against PTPO-Shared are not significant under root-level inference. The same-batch graph and degree-corrected pair implementations agree to a maximum parameter difference of $3.2 \times 10^{-7}$, and Typed-Pair with independent sampling consumes 5.94B generation tokens versus PTPO's 4.82B: at a matched 4.82B budget its Switch scores are 29.13/24.36 against PTPO's 35.94/30.91 (Tables S13 and S8). This supports rollout reuse as a budget advantage, not a mathematically stronger objective.

## 7 Limitations

ECR evaluates observable evidence use, not the causal faithfulness of latent reasoning. Its scope is the audited evidence universe and certificate space; fixed eligibility prevents method-selected denominators but does not remove proof-audit errors. The coordinate-wise evaluation also matters substantively: PTPO's supported Switch advantage does not imply significant gains in Change or Preserve over CRPO-Evidence. Endpoint matching controls measured aggregate scores, not training history or the distribution of instance-level errors. Transfer claims concern RL-held-out relations under shared Stage-A supervision and the reported static or agent interface, not unrestricted zero-exposure generalization. The repeated-world control remains nonzero even under greedy decoding (Table S7); builder-level holdouts, reverse-direction transfer, and broader external evaluation remain untested.

## 8 Conclusion

Language models should revise the conclusions warranted by an evidence change, not merely produce a different answer. EviPair makes this distinction measurable through withdrawal, preservation, and proof adaptation on fixed evaluation objects. Endpoint reinforcement improves answering while worsening preservation, and similar aggregate endpoints retain different response profiles. Structured intervention training improves RL-held-out switching, with the clearest comparative gain in alternative-proof adaptation. The resulting contribution is a behavioral evaluation target and evidence about its training transfer, rather than a higher-order optimization objective.

## AI USE STATEMENT

AI assistance was used only for typesetting and language polishing of the manuscript. The evaluation objects, eligibility rules, atom mappings, exposure ledger, and interface assignments reported here are stated in Appendix A and are fixed before any evaluated model runs.

## REPRODUCIBILITY STATEMENT

The evaluation objects, eligibility rules, aggregation conventions, reward definitions, and optimization assumptions are specified in Sections 3–5 and Appendices A–B. Tables S2 and S1 document dataset scale and frozen settings; Tables S10, S13, and S18 provide seed-level summaries, recorded implementation checks, and exposure information. Figure source files and the data used for the endpoint-response plot accompany the manuscript project. This manuscript bundle is not an experimental-code or raw-trajectory release.

## REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE style optimization for learning from human feedback in LLMs, 2024. URL https://arxiv.org/abs/2402.14740.

Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. QAMPARI: A benchmark for open-domain questions with many answers. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 97–110. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.gem-1.9/.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Dazhao Du, Jian Liu, Jialong Qin, Tao Han, Bohai Gu, Fangqi Zhu, Yujia Zhang, Eric Liu, Xi Chen, and Song Guo. Learning spatiotemporal sensitivity in video LLMs via counterfactual reinforcement learning, 2026. URL https://arxiv.org/abs/2605.21988.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients, 2017. URL https://arxiv.org/abs/1705.08926.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Wangzhen Guo, Qinkang Gong, Yanghui Rao, and Hanjiang Lai. Counterfactual multihop QA: A cause-effect approach for reducing disconnected reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4214–4226. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.231. URL https://aclanthology.org/2023.acl-long.231/.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.09516.

Kyungjae Lee, Seung-won Hwang, Sang-eun Han, and Dohyeon Lee. Robustifying multi-hop QA through pseudo-evidentiality training. In *Proceedings of the 59th Annual Meeting of the Association*

*for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6110–6119. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.476. URL https://aclanthology.org/2021.acl-long.476/.

Chenyu Lin, Yilin Wen, Du Su, Hexiang Tan, Fei Sun, Muhan Chen, Chenfu Bao, and Zhonghou Lv. Resisting contextual interference in RAG via parametric-knowledge reinforcement. In *International Conference on Learning Representations*, pp. 46721–46742, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4dbcd4ba82866fdefc7ea0d15e9a72f2-Paper-Conference.pdf.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of NAACL-HLT, Volume 1 (Long Papers)*, pp. 809–819. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Jungmin Yun, June Hyoung Kwon, and Youngbin Kim. CRiT-QA: Evaluating multi-hop reasoning with counterfactual chains and distractor traps. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 5246–5255. ELRA Language Resource Association, 2026. doi: 10.63317/2jvxj7kwecuo. URL https://aclanthology.org/2026.lrec-1.410/.

JunShuo Zhang, Chengrui Huang, Feng Guo, Zihan Li, Ke Shi, Menghua Jiang, Jiguo Yu, Shuo Shang, and Shen Gao. DPEPO: Diverse parallel exploration policy optimization for LLM-based agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46367–46389. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2151. URL https://aclanthology.org/2026.acl-long.2151/.

# A Detailed Evaluation and Training Protocol

## A.1 Independent worlds and the verifier boundary

A root problem $r$ consists of a query $q$ and a finite, audited evidence universe. A world $w_v$ specifies an available evidence set $E_v$ and its tool-visibility configuration. Each node is evaluated in a fresh session, with an independent sampling stream when decoding is stochastic. The policy receives only the query and observations available in that world. It receives neither another world's answer nor the intervention label, gold set, support graph, or verifier metadata. These restrictions apply to prompts, retrieval results, caches, summaries, and error messages. Shared evidence keeps its identifier across worlds; deleted items are not replaced by renumbering the survivors.

An output $o_v = (Y_v, C_v)$ consists of a normalized semantic set and a certificate for each predicted atom. Throughout this paper, *revision* and *switching* denote relations between independently generated outputs. They do not denote a sequential editing task in which a model reads and corrects its previous answer. This operational definition measures observable evidence-dependent behavior, not the internal causal faithfulness of a reasoning trace.

**Task representations.** For MuSiQue-Full (Trivedi et al., 2022) the native fields are `answerable`, `answer`, and `support_ids`. An unanswerable world has the empty semantic set. An answerable world contributes one atom per conclusion that its audited evidence independently supports, so a world whose audited decomposition yields several sub-conclusions carries several atoms. For FEVER (Thorne et al., 2018) the native fields are `label` and `evidence_ids`; an atom is one claim–label judgment, and a derived world covers a group of claims that share an audited evidence universe, so it carries one atom for each claim whose gold label in that world is SUPPORTS or REFUTES and none for a claim labeled NEI. A world whose claims are all NEI therefore carries the empty semantic set. As for MuSiQue-Full, the number of atoms in a FEVER world is fixed by the audited construction rather than by the count of native records. QAMPARI (Amouyal et al., 2023) uses a set of normalized answer entities, each with its own `evidence_ids`.

Because a world's gold atom set may contain several atoms, one edge may carry several atomic obligations at once: a selective set-removal edge requires withdrawal of $D_e^*$ and preservation of $P_e^*$ on the same edge, and a Forced Switch edge may designate several targets $K_e$. The number of atoms per world and per edge is a property of the audited construction and is reported in Tables S4 and S22; it is not inferred from the native output fields. This holds on every task: a FEVER world can lose support for one of its claims while retaining another, and a Forced Switch edge can designate several targets even though each individual judgment is binary.

Entity aliases and text normalization are frozen before evaluation. Duplicate entities are merged, with the union of their citations; unknown entities remain erroneous predictions rather than being discarded. Correct unanswerable and NEI outputs require empty answer content and empty citations. A malformed output is the distinct state $\perp$, never a legitimate empty set. Pure answer-list ordering does not affect a set-valued score.

**Worked examples.** Two minimal, fictional instances fix the mapping from native fields to atoms, certificates, and node reward (Appendix A.6 gives the general formulas). For MuSiQue-Full, a query with two independently sufficient sub-conclusions (say, a birth year and a nationality) yields an answerable world with gold atoms $A_v = \{a_1, a_2\}$. An output predicting only $Y_v = \{a_1\}$, correctly, with a valid certificate for $a_1$, gives answer F1 $= 2|Y_v \cap A_v|/(|Y_v| + |A_v|) = 2/3$, so $m_v = 2/3$; since the one predicted atom is certified, $e_v = m_v$, and $R_v = f_v\, m_v = \frac{2}{3} f_v$. For FEVER, a world covering two audited claims carries gold atoms $A_v = \{a_1, a_2\}$, where $a_1 = (x_1, \textsc{Supports})$ and $a_2 = (x_2, \textsc{Refutes})$. An output judging both claims correctly and certifying both with audited evidence gives $Y_v = A_v$, hence $\ell_v = 1$, $c_v = 1$, and $R_v = f_v$. An output judging $x_1$ correctly but abstaining on $x_2$ gives $Y_v = \{a_1\}$, so $\ell_v = 1/2$; with $a_1$ certified, $c_v = 1$ and $R_v = f_v(\frac{1}{2} + \frac{1}{2})/2 = f_v/2$. Predicting the opposite label for a claim keeps that atom outside $Y_v$ and lowers $\ell_v$ accordingly, regardless of certificate validity. A selective set-removal edge is the same arithmetic read across two worlds: one atom of $A_u$ drops out of the target gold set and enters $D_e^*$ while another stays in $P_e^*$.

### A.2 SUFFICIENT SETS AND CERTIFICATES

Let $\mathcal{S}_r^{\mathrm{raw}}(a)$ be the audited family of sufficient evidence sets for atom $a$. Before freezing eligibility, remove redundant supersets:

$$\mathcal{S}_r(a) = \{S \in \mathcal{S}_r^{\mathrm{raw}}(a) : \nexists S' \in \mathcal{S}_r^{\mathrm{raw}}(a) \text{ with } S' \subsetneq S\}. \tag{6}$$

This is inclusion minimization within the audited family, not an assertion that annotation alone exhausts every proof in an unrestricted corpus. A proof may require several jointly sufficient items; individual members need not be sufficient. Since the family is finite, removing its redundant supersets preserves whether it contains a proof inside any available evidence set. In particular,

$$\mathrm{Supported}(a, w_v) \iff \exists S \in \mathcal{S}_r(a) : S \subseteq E_v. \tag{7}$$

A missing annotation is not by itself evidence of absent support. Completeness of the relevant proof search is assessed within the declared finite evidence and certificate space.

Let $O_v(\tau_v) \subseteq E_v$ be the evidence that the trajectory is allowed to cite. For direct-context evaluation this is the supplied evidence; for the retrieval agent it is the content opened in that trajectory. For a nonempty semantic atom, certificate validity is

$$\mathrm{Valid}_v(a, C; \tau_v) = \mathbf{1}\{a \in A_v,\ C \subseteq O_v(\tau_v),\ \exists S \in \mathcal{S}_r(a) : S \subseteq C\}, \tag{8}$$

where $A_v$ is the audited gold semantic set. Every cited item must be available and observable, and the certificate must contain a complete proof. Extra available but irrelevant citations need not destroy sufficiency; their usefulness is a separate citation-precision issue. An unavailable or unobserved citation invalidates the certificate. The legal empty-output convention for unanswerable/NEI worlds is handled separately, not by inventing a proof of a nonexistent answer.

Minimization matters for adaptation: the disappearance of an irrelevant item from a redundant superset must not be interpreted as destruction of a proof when a smaller sufficient subset survives. Necessary removal must hit every audited sufficient set for the affected atom. Such a deletion set is called *minimal* only if its minimality has itself been established.

### A.3 FORCED SWITCHING VERSUS NATURAL REDUNDANCY

For $e = (u, v)$ define

$$D_e^* = A_u \setminus A_v, \quad P_e^* = A_u \cap A_v, \quad A_e^+ = A_v \setminus A_u, \qquad \mathcal{S}_u(a) = \{S \in \mathcal{S}_r(a) : S \subseteq E_u\}. \tag{9}$$

All target sets and eligibility predicates are frozen before any evaluated model runs.

**Fixed Forced Switch eligibility.** An eligible Forced Switch edge has $A_u = A_v$ and a nonempty fixed target set $K_e \subseteq P_e^*$. For every $a \in K_e$, the source contains at least one sufficient proof, the target contains at least one sufficient proof, and

$$\text{for every } S \in \mathcal{S}_u(a), \qquad S \not\subseteq E_v. \tag{10}$$

Thus no complete source proof remains sufficient in the target, regardless of which valid certificate a model selected. Shared bridge facts are permitted if they do not constitute a surviving complete proof. Relabeling identical text, duplicating a passage, or merely changing an identifier does not establish eligibility. Candidates whose old proofs cannot all be confirmed invalid are excluded from the Forced Switch eligibility set before freezing.

This is generally a support-replacement intervention, not pure deletion. If $E_v \subseteq E_u$, every sufficient target proof would also be a complete source proof surviving in $E_v$, contradicting (10). An eligible switch therefore requires a genuine change beyond deleting evidence, such as introducing alternative support. The policy still sees only its own world's resulting evidence.

For a fixed target define

$$h_e(a) = \mathbf{1}\{a \in Y_u \cap Y_v\}\, \mathrm{Valid}_u(a, C_u(a); \tau_u)\, \mathrm{Valid}_v(a, C_v(a); \tau_v), \tag{11}$$

and set it to zero on parse or tool failure. Target validity rules out citations to deleted sources. With $f_v$ the valid-format indicator, the edge success indicator is

$$I_e^{\mathrm{sw}} = f_u f_v \mathbf{1}\{Y_u \subseteq A_u,\ Y_v \subseteq A_v\} \prod_{a \in K_e} h_e(a). \tag{12}$$

SwitchAcc is $\sum_{e\in\mathcal{E}_{\mathrm{sw}}} I_e^{\mathrm{sw}}/|\mathcal{E}_{\mathrm{sw}}|$, using the same fixed edge set for all methods. Source errors, missing citations, invalid certificates, and target failures all remain in this denominator. A correct non-target atom omitted by a model is penalized by Preserve and Endpoint; switch success alone is not full task success. Multiple valid target proofs are accepted.

The unit-level diagnostic instead averages each edge's per-target success fraction with equal edge weight over the same frozen edge set,

$$U_{\mathrm{edge}} = \frac{1}{|\mathcal{E}_{\mathrm{sw}}|} \sum_{e\in\mathcal{E}_{\mathrm{sw}}} \frac{1}{|K_e|} \sum_{a\in K_e} h_e(a). \tag{13}$$

It keeps every eligible edge in the denominator; only the aggregation order differs from SwitchAcc, since $|K_e|$ need not be constant across edges. Because an all-succeed indicator never exceeds that edge's own per-target average, $I_e^{\mathrm{sw}} \le \frac{1}{|K_e|}\sum_{a\in K_e} h_e(a)$ on every edge, so the edge-level conjunction is necessarily no larger than this diagnostic. Table S22 reports $U_{\mathrm{edge}}$; a root-balanced switch summary is likewise a separate aggregation, not a replacement for the fixed-edge ratio. The distance between the two aggregations has two sources. One is within-edge dependence rather than target count alone: targets on one edge share a world pair and a single policy, so their successes are strongly correlated and the conjunction penalty need not grow proportionally with $|K_e|$, which is why FEVER and QAMPARI carry different mean target counts yet similar gaps. The other is the output-level gate $\mathbf{1}\{Y_u \subseteq A_u,\ Y_v \subseteq A_v\}$: edge success also fails when either endpoint carries an extra incorrect semantic prediction, while $U_{\mathrm{edge}}$ never applies that gate.

**Natural redundancy.** A natural-redundancy intervention deletes evidence while retaining at least one proof for an affected-but-preserved atom. Its fixed target set is

$$N_e = \{a \in P_e^* : \exists S \in \mathcal{S}_u(a),\ S \not\subseteq E_v, \text{ and } \exists S' \in \mathcal{S}_u(a),\ S' \subseteq E_v\}. \tag{14}$$

SupportAdaptAcc is the mean of $\prod_{a\in N_e} h_e(a)$ over all natural-redundancy edges with $N_e \neq \emptyset$, with failures scored zero. A corresponding unit rate uses the fixed total number of targets. If the source output already uses a surviving proof, retaining it is correct: no certificate change is demanded solely because some other proof was removed. The fixed-edge result is reported in Table S9(a).

For the conditional diagnostic, an actual trigger is a target in $N_e$ whose source semantics and certificate are correct, but whose surviving certificate content $C_u(a) \cap E_v$ contains no complete sufficient proof. SwitchCond measures target success on these triggered units, with trigger coverage equal to their number divided by $\sum_e |N_e|$. If no unit triggers, the conditional rate is not applicable. Because trigger sets depend on model outputs, their conditional rates are descriptive, not a common-denominator ranking (Table S9(b)). They are never substituted for SwitchAcc.

### A.4 Set metrics, empty sets, and aggregation

For a prediction set $X$ and reference set $T$, let $\mathrm{TP} = |X \cap T|$. When both are nonempty, precision is $\mathrm{TP}/|X|$, recall is $\mathrm{TP}/|T|$, and F1 is $2\mathrm{TP}/(|X| + |T|)$. When exactly one set is empty all three scores are zero; when both are empty all three are one. The latter convention never gives a free main-coordinate score, since Change and Preserve have nonempty-reference eligibility requirements.

For successfully parsed endpoints, define

$$\widehat{D}_e = Y_u \setminus Y_v, \quad \widehat{P}_e = Y_u \cap Y_v, \quad \widehat{A}_e = Y_v \setminus Y_u. \tag{15}$$

Change compares $\widehat{D}_e$ to $D_e^*$ only when $D_e^* \neq \emptyset$; Preserve compares $\widehat{P}_e$ to $P_e^*$ only when $P_e^* \neq \emptyset$. Either endpoint's parse failure makes each applicable main score zero. Constant empty answers have zero Preserve recall; constant wrong answers are false positives. A selective set-removal edge can contribute to both coordinates. An ineligible coordinate is marked not applicable, not assigned zero or one. Precision and recall follow the same aggregation as F1 (Table S21); averaging F1 is not equivalent to taking the harmonic mean of averaged precision and recall.

For precision, recall, or F1 separately, let $\mathcal{E}_{r,t}^m$ be the fixed eligible edges of root $r$ and primary type $t$ for coordinate $m$. Let $\mathcal{R}_t^m$ be the roots with nonempty $\mathcal{E}_{r,t}^m$, and $T^m$ the applicable types. The dataset score is

$$M_m = \frac{1}{|T^m|} \sum_{t\in T^m} \frac{1}{|\mathcal{R}_t^m|} \sum_{r\in\mathcal{R}_t^m} \frac{1}{|\mathcal{E}_{r,t}^m|} \sum_{e\in\mathcal{E}_{r,t}^m} m_e. \tag{16}$$

This averages edges within a root/type, roots within a type, and finally types. Eligibility is independent of model success. Raw eligible-edge counts describe coverage, not uniform edge weights for these two coordinates. SwitchAcc uses its literal fixed-edge ratio instead. Cross-task macro averages use all three datasets for Change/Preserve and only FEVER and QAMPARI for Switch. There is no single ECR aggregate ranking. MuSiQue-Full Switch is not applicable because no audited eligible alternative-proof family was obtained.

Semantic Endpoint is distinct from evidence validity. MuSiQue-Full uses answer F1 on sufficient nodes, averaging these nodes within roots and then roots. Sufficient- and insufficient-node answerability are reported separately. FEVER uses atomic label recall on derived worlds: a world scores the fraction of its gold atoms whose claim–label judgment the output reproduces, and deduplicated nodes are averaged within roots and then roots. This is a gold-atom recall rather than a three-way accuracy over all claims: claims whose gold label is NEI lie outside the denominator, and extra predicted judgments outside $A_v$ are not charged here, since they enter TargetError and UnexpectedAdd instead. QAMPARI uses answer-set F1 against current shard-bounded gold, averaged over deduplicated nodes within roots and then roots. Node degree never supplies endpoint weights. Native evidence-aware task scores retain their own names (Table S20), and a cross-task Endpoint mean mixes scales rather than defining a primary ranking.

**Additional error channels.** On parsed outputs, the diagnostic quantities include

$$\mathrm{UnexpectedAdd}_e = \frac{|\widehat{A}_e \setminus A_e^+|}{\max(1, |\widehat{A}_e|)}, \tag{17}$$

$$\mathrm{NewUnsupported}_e = \frac{|\widehat{A}_e \setminus A_v|}{\max(1, |Y_v|)}, \tag{18}$$

$$\mathrm{TargetError}_e = \frac{|Y_v \setminus A_v|}{\max(1, |Y_v|)}. \tag{19}$$

UnexpectedAdd is not synonymous with hallucination: it also includes newly supplied, previously omitted members of $P_e^*$. The recovered-content count is $|\widehat{A}_e \cap P_e^*|$; its nonzero-event frequency is the edge-incidence version reported as RecoveredPreserved in Table S6. On $D_e^* = \emptyset$ edges, AnyRemoval is $\mathbf{1}\{\widehat{D}_e \neq \emptyset\}$ and FalseDrop is $\mathbf{1}\{\widehat{D}_e \cap P_e^* \neq \emptyset\}$. Removing an unsupported source prediction, measured by $|\widehat{D}_e \setminus A_u|$ or its nonzero-event rate, is not a false retraction. Count-valued and incidence-valued summaries should not be interchanged. The $P_e^* = \emptyset$ layer is reported separately with Preserve not applicable.

Parsed-output diagnostics carry their coverage and do not silently drop formatting failures from the main metrics. Reporting the $D^* = \emptyset$-layer union of format failure with FalseDrop prevents a model from appearing to reduce semantic errors merely by producing malformed outputs; it is a layer quantity, and the corresponding all-edge bound is stated with Table S6 so that it is not read as an all-edge rate. T-Exact requires complete semantic and certificate correctness at both endpoints and all applicable constraints; it is an auxiliary conjunction (Table S5), not independent evidence of a new capability. Source-correct conditional analysis (Table S12) likewise accompanies, rather than replaces, the full-denominator evaluation.

### A.5 Data construction and audit scope

Each root record specifies the original example, dataset, split, query, evidence-universe version, constructor version, and audit status. World records contain stable evidence identifiers and text hashes, visibility settings, gold semantics, and sufficient-set families. Edge records specify direction, primary type, intervention family, removed/added identifiers, $D_e^*, P_e^*, A_e^+$, switch and natural-adaptation targets, certificate eligibility, and normalized weights. Identical evidence/visibility worlds are deduplicated; self-loops are disallowed. Reverse edges require their own verification and are not additional independent observations.

Splitting precedes world derivation: all worlds of a root remain in one split. The leakage audit treats question paraphrases and restatements of an atomic fact as related units and checks document, fact-family, text, and replacement-template overlap. The log records document/fact-family isolation

controls; this does not establish absence of pretraining contamination. The constructor, alternative-evidence selection, audit decisions, and root/edge manifests are frozen, with content hashes, before running evaluated models. Construction failures may be excluded before this freeze; model failures after it are retained.

MuSiQue-Full uses audited sufficient/insufficient transformations and irrelevant-evidence controls without assuming every native transformation is pure deletion. FEVER uses finite sentence worlds, including derived NEI, and resolves or excludes conflicting support/refutation. QAMPARI gold is recomputed relative to the bounded evidence shard rather than inherited unchanged from an unrestricted corpus. Candidate shards start with 24–32 passages; a candidate that cannot retain a complete sufficient proof within the construction constraints is discarded rather than truncated with stale gold. The supplied records do not specify a numerical final passage cap beyond this candidate range, so no additional cap is inferred here. Worlds share one underlying evidence universe instead of independently resampling negatives.

Rule checks cover set operations and sufficient-family inclusion minimization. Independent semantic auditing checks omitted proofs, joint cross-passage support, contradictory evidence, and answer associations. The recorded training-root funnels are 4,850 to 4,216 for MuSiQue-Full, 5,600 to 5,040 for FEVER, and 3,200 to 2,880 for QAMPARI. Training uses high-confidence rules with stratified manual audits of 12%, 12%, and 15%, respectively; the recorded semantic disagreement rates are 4.60%, 3.80%, and 5.20%. Confirmatory Forced Switch and critical edges receive full review, with two independent judgments and third-person adjudication. Among exclusions, failure to establish invalidation of all old proofs accounts for 41.20%, unresolved joint support for 24.60%, unresolved contradiction for 18.30%, and insufficient alternative support for 15.90%. These are construction/audit statistics, not model-performance scores.

Table S2 records the resulting graph sizes: 12,136 training roots, 41,098 worlds, and 36,243 typed edges; the test collection has 1,548 roots, 5,241 worlds, and 4,622 edges. Table S4 records eligibility coverage, including 1,971 Change edges, 3,185 Preserve edges, and 485 Forced Switch edges. These coordinate counts overlap and do not partition the graph. Evidence position, source order, length, identifier stability, and edit volume are controlled rather than treated as proof of a support change. No-op and matched-volume sham edits provide supporting diagnostics (Table S7); their existence does not by itself exclude every constructor shortcut.

**Training exposure and generalization.** The main curriculum contains necessary/selective removal, irrelevant preservation, and natural redundancy, with no Forced Switch training edges (Table S4). Natural redundancy nevertheless exposes alternative-support adaptation, so this curriculum is not described as having no alternative-support supervision. The stricter condition removes natural-redundancy edges together with the dedicated alternative-support nodes built for them; because the two are removed as one package, the resulting difference is not attributed to the reward component alone.

Every holdout is applied at the root level: a held-out root contributes no world and no edge to the RL graph, and none of its worlds is retained (Table S18). The Forced Switch relation is held out in full, so its training count is zero by construction rather than because the constructor never produced such edges. Topology holdouts concern joint multi-sentence proofs, dependency hop at least four, and the per-target predicate of at least three alternative proofs; these predicates are fixed before Stage-A and apply to roots throughout. Stage-A itself uses single-world interface supervision over all training roots and is shared by every method within a seed, so the holdouts act on the RL graph only. The exposure ledger, rather than the absence of a reward component alone, determines the interpretation.

The reported cross-dataset direction is MuSiQue to QAMPARI under a unified direct-context interface, with target-task examples excluded from training and selection. Retrieval-agent transfer is a separate interface condition (Table S14). The second backbone is a replication across model families, not a substitute for cross-task transfer. Constructor/template-level holdouts, the reverse transfer direction, and an external CRiT-QA transfer evaluation are not part of the reported scope. Revise, involving audited new semantic gold, is an extension of the formulation and not an additional completed intervention family.

### A.6 NODE REWARDS AND TYPED EDGE REWARDS

Node rewards score one world's output, while edge rewards emphasize its relation to another independent output. They share semantic matching and certificate verification with evaluation but are not identical to the main metrics.

For MuSiQue-Full let $m_v$ be answerability correctness times answer F1 on sufficient worlds, or answerability correctness on insufficient worlds. Let $e_v = m_v$ times valid-certificate status, with a correct insufficient empty output assigned valid evidence status. Then $R_v = f_v(m_v + e_v)/2$. For FEVER, let $\ell_v$ be the fraction of the world's gold atoms that the output reproduces and $c_v$ the fraction of reproduced atoms whose certificates are valid; the legal empty output against an empty gold set scores $\ell_v = c_v = 1$, and a nonempty output against empty gold scores zero, $R_v = f_v(\ell_v + \ell_v c_v)/2$. For QAMPARI, define

$$\mathrm{GF1}_v = \frac{2|\{a \in Y_v \cap A_v : \mathrm{Valid}_v(a, C_v(a); \tau_v) = 1\}|}{|Y_v| + |A_v|}, \qquad R_v = \frac{f_v}{2}\big(\mathrm{F1}(Y_v, A_v) + \mathrm{GF1}_v\big), \tag{20}$$

with both F1 terms equal to one for a legal double-empty output. An invalid certificate does not remove its prediction from the grounded-F1 denominator. All rewards lie in $[0, 1]$; parse failures receive zero.

Let $c_e$ denote applicable Change-F1 and $p_e$ applicable Preserve-F1. The Forced Switch component $s_e$ is the mean of $h_e(a)$ over the fixed targets $K_e$; the natural-adaptation component $g_e$ is its mean over $N_e$, with no unnecessary requirement to change a surviving proof. Let $u_v = |Y_v \cap A_v|/|Y_v|$ for nonempty predictions, zero for empty predictions against nonempty gold, and one for the legal double-empty case. The typed edge reward is

$$R_e = f_u f_v u_u u_v \, \frac{1}{|\mathcal{C}_e|} \sum_{x \in \mathcal{C}_e} x. \tag{21}$$

The frozen component set $\mathcal{C}_e$ contains $c_e$ and, when applicable, $p_e$ for Drop; $p_e$ for irrelevant Preserve; $p_e, g_e$ for natural redundancy; and $p_e, s_e$ for Forced Switch. This last rule defines the evaluator/reward interface even though Forced Switch edges are absent from the training graph. Edges without any applicable component are excluded from the primary training graph. Removing a component in an ablation re-averages the remaining components; if none remain, the edge reward is zero and its original $b_e$ is retained. Evaluation eligibility never changes with a reward ablation or with model success.

The graph objective and incident-node credit are formalized in Appendix B. Endpoint rewards occur exactly once in the node term. Repeating them uncorrected on every neighboring edge would overweight high-degree worlds. Independent typed-pair sampling, graph reuse, and random matching are implementations to be distinguished from the chosen typed reward. The degree-corrected pair decomposition, including isolated nodes and importance weights, is given in Appendix B.4.

**CRPO-Evidence baseline kernel.** CRPO-Evidence replaces the typed edge return $R_e$ with a binary semantic-relation kernel while keeping every other term (endpoint rewards, $\lambda$, $a_v$, $b_e$, and the node/edge graph) identical to PTPO. For edge $e = (u, v)$ and rollout indices $i, j$, let $c_i = \mathbf{1}\{Y_u^i = A_u\}$ be source correctness; let $h_j = \mathbf{1}\{Y_v^j \neq A_u\}$ when the edge's primary relation requires semantic change (Drop) and $h_j = \mathbf{1}\{Y_v^j = A_u\}$ when it requires retention (Preserve or Switch, whose gold set is unchanged). Format failures set $c_i$ or $h_j$ to zero directly, and QAMPARI compares normalized answer sets rather than single labels. The kernel

$$K_{ij}^e = f_{ui} f_{vj} \, c_i h_j \tag{22}$$

gates the same row/column averaging PTPO applies to $Q_{ij}^e$ (Eq. (4)). Its reported credit convention, *joint standardization*, centers each node's credit against every node of the same root graph and divides by their population standard deviation plus a pre-registered $\epsilon$, rather than PTPO's node-local leave-one-out baseline; *Binary + Shared* and $\rightarrow$ *typed kernel* in Table S26 isolate the kernel and credit choices separately. This is a graph-adapted reproduction of CRPO's reported mechanism, not a claim of an unmodified original implementation.

### A.7 TRAINING CONFIGURATION AND TOOL INTERFACE

The primary backbone is Qwen3-4B-Base; Llama-3.1-8B-Instruct is the confirmatory second family. Stage-A uses single-world interface supervision for two epochs, learning rate $10^{-5}$, cosine scheduling, batch size 64, and maximum sequence length 2,048. It does not provide paired relation labels or the other endpoint's output. Methods with matched exposure share the same Stage-A checkpoint within each seed.

The frozen RL settings are learning rate $10^{-6}$, 2,000 updates, a root batch of eight roots per micro-batch with sixteen micro-batches accumulated per update, $G = 4$ rollouts per node, $\lambda = 0.50$, clipping parameter $\epsilon = 0.20$, and KL coefficient zero (Table S1). Advantages use node-level leave-one-out centering with no per-group standard-deviation normalization. Rewards and advantages are detached. Losses sum policy-token contributions, mask prompts, tool observations, and padding, average over the $G$ samples and root batch, and do not divide by the number of nodes again. Zero-advantage samples are not regenerated and unsuccessful samples are not removed. Training temperature is 1.00 and maximum generation length is 1,024 tokens per policy call. A MuSiQue-Full or FEVER rollout issues one such call; a QAMPARI rollout is the full `search`/`open`/`final` trajectory of up to six calls, so its total generated length can exceed 1,024 tokens even though no single call does. Main evaluation is greedy; the separate temperature-1.00 comparison is in Table S24.

The QAMPARI agent exposes `search`, `open`, and `final`, with a budget of at most six calls under the recorded interface. Search provides navigation identifiers and non-certifying clues; final citations must refer to content opened in that trajectory. An available proof that the agent fails to retrieve remains in gold, so retrieval failure is not relabeled as an unanswerable world. Malformed actions, timeouts, and budget failures remain in the evaluation denominator. The verifier timeout is 30 seconds. The supplied configuration does not give a numerical retry count, and we do not infer one or an undocumented action schema.

Runs use one node with four NVIDIA A40 GPUs (48 GB each) and 1 TB of host memory. The logged stack is Ubuntu 22.04.4 LTS, Python 3.11.9, PyTorch 2.4.1 with CUDA 12.4, Transformers 4.44.2, vLLM 0.6.1.post2 with prefix caching, FlashAttention-2 2.6.3, DeepSpeed 0.15.1 with ZeRO-3 and no offload, and bf16 precision; retrieval uses Pyserini 0.24.0 (BM25) and FAISS 1.8.0. Cache reuse does not relax the world-visibility boundary. These are the versions recorded for the runs, not a claim of compatibility established by this manuscript.

Costs in Table S8 distinguish node rollouts, generation tokens, edge-verifier calls, and device time. Its wall-clock column is the GPU-hour total divided by four, not an independent additional measurement. All-pair reward combinations are not additional independent generations. The matched-generation-budget typed-pair control has a different effective update count, so it is not described as matching every compute dimension simultaneously.

### A.8 CHECKPOINT SELECTION AND STATISTICAL REPORTING

Checkpoints are saved every 50 RL steps, giving 40 candidates. Matching is performed independently for each seed on development data within a predeclared common endpoint range. MuSiQue-Full simultaneously matches sufficient-node answer F1 within $\pm 0.50$ points and sufficient-/insufficient-node answerability within $\pm 1.00$ points each. FEVER label recall and QAMPARI answer F1 use $\pm 0.50$ points. Ties prefer fewer training tokens and then the fixed checkpoint identifier, not a better ECR score. After selection, evaluation uses the complete frozen test graphs, without test-based checkpoint selection or favorable subsets. The reported response curves use six pre-registered checkpoints, not all 40 saved candidates. Table S23 records the selected checkpoints and their development metrics.

Three thresholds serve distinct purposes. $\delta_{\mathrm{match}}$ selects a checkpoint on development data. The cross-method threshold $\kappa = 1.50$ points determines whether the resulting test endpoint range remains comparable. The non-inferiority margin $\eta = 1.00$ points is assessed against the lower confidence bound of an endpoint difference: non-inferiority requires that bound to exceed $-1.00$. A nonsignificant superiority test does not establish non-inferiority. If endpoint comparability fails, the comparison is reported as unmatched rather than rescued by widening the tolerance or inventing interpolated test scores. Even a valid match controls only the measured endpoint summaries, not optimization history or the complete instance-level error distribution.

The five training seeds are 17, 42, 73, 101, and 2024. Main-table $\pm$ entries are sample standard deviations across these seeds, computed separately for each method, dataset, and coordinate. They describe training reproducibility, not inferential confidence intervals. Base is evaluated once with fixed prompting and carries no training-seed spread. Per-seed outcomes and contrasts are in Tables S10 and S15; seed variation is reported separately from root sampling.

Macro means, the $\Delta$ rows of Table 1, and the difference block of Table S27 are computed from unrounded per-seed values and rounded once for display. Subtracting the displayed two-decimal entries can therefore differ from a reported difference by 0.01; this is display rounding, not a second aggregation.

Inference uses paired root-clustered bootstrap comparisons, with 10,000 root resamples per seed in the recorded table protocol. A resampled root carries all of its worlds and edges, with the same sampled roots used for the two methods. Each replicate recomputes the declared aggregation; for SwitchAcc this includes recomputing the success numerator and eligible-edge denominator. The sampling unit is not an individual edge or an entry of a $G \times G$ reward matrix. The intervals and operating characteristics are in Tables S3 and S16. Clustering inflates variance by a design effect of 1.41–1.66 across coordinates, leaving an edge-level effective sample size of only 60–71% of the raw denominator (Table S16). This inflation is measured against independent edge-level resampling, not against the five-seed spread, which tracks a different source of variation; root-clustered intervals are therefore markedly wider than the seed standard deviations in the main tables, and the two must not be substituted for each other. Visual overlap or separation of seed standard deviations supplies no significance decision.

Switch is the sole pre-registered primary testing family; its six PTPO-versus-baseline contrasts use Holm correction. Change, Preserve, and Endpoint are exploratory coordinates with uncorrected $p$-values, not additional corrected confirmatory families (Table S17). Absolute performance and difference intervals accompany the tests. Reported $p$-values are two-sided Wald values obtained from the same root-clustered bootstrap standard errors that set the interval widths in Table S3, evaluated under the normal approximation to the studentized difference. They are consequently not empirical tail frequencies: 10,000 resamples could not resolve probabilities below $10^{-4}$, whereas a normal tail is defined at any magnitude. Each interval and its $p$-value are two summaries of one studentized quantity rather than independent evidence, and the interval bounds, not the $p$-value, carry the magnitude information.

Finally, Table S19 reports an *implied*, not directly measured, correlation. Its calculation recovers a normal-equivalent standard error from an interval half-width and uses

$$\mathrm{SE}_{\mathrm{diff}} = \mathrm{SE}_{\mathrm{single}} \sqrt{2(1-\rho)}. \tag{23}$$

This inversion uses a common marginal-variance approximation; it is not the unrestricted covariance identity when the two methods have different marginal variances. The derived quantity describes the reported interval scales and does not enter significance decisions or establish that matching removed all confounding.

# B Properties and Limits of Edge-Additive Policy Credit

This section establishes the gradient identity used by PTPO, a conditional-averaging result for the raw matching estimator, and a finite non-identifiability example for aggregate endpoint summaries. These are properties of the stated objective and observation model, not claims of a new policy-gradient identity or of a higher-order graph objective.

## B.1 Sampling assumptions and normalized objective

Fix a root $r$ and its finite directed graph $(V_r, \mathcal{A}_r)$, with no self-loops. Parallel or reverse edges, if present, are separately audited and counted in the weights. Let $p_{\theta,v}(\tau)$ be the trajectory distribution in world $w_v$. Conditional on the root and its fixed worlds, the joint policy is the product

$$p_\theta((\tau_v)_{v \in V_r} \mid r) = \prod_{v \in V_r} p_{\theta,v}(\tau_v). \tag{24}$$

The parameters may be shared across worlds; independence concerns trajectory sampling, not parameter blocks. Within each node, $G \geq 2$ trajectories are independent and identically distributed, and the sample arrays of distinct nodes are independent. Environment transitions and tool observations have no direct dependence on $\theta$ conditional on actions. Thus the trajectory score

$$z_v(\tau) = \nabla_\theta \log p_{\theta,v}(\tau) = \sum_{t \in \mathcal{T}(\tau)} \nabla_\theta \log \pi_\theta(x_t \mid h_t, w_v) \tag{25}$$

contains only policy-generated tokens. Prompt, observation, and padding tokens are excluded.

Assume parameter-independent support locally in $\theta$, differentiability, and sufficient integrability to interchange differentiation and expectation. In particular, $\mathbb{E}[z_v] = 0$. Rewards are measurable functions of their stated trajectories and frozen verifier metadata, without direct parameter dependence. Weights and the root distribution are fixed independently of policy outputs. For the covariance statement below, also assume finite second moments of the score-weighted returns.

For a graph with at least one applicable edge, define

$$a_v = \frac{1}{|V_r|}, \qquad b_e = \frac{1}{|T_r| n_{\rho_e}}, \qquad \sum_v a_v = \sum_e b_e = 1, \tag{26}$$

where $T_r$ is the set of applicable primary relation types and $n_t$ their edge counts. Consequently, the sum of $b_e$ within a single type is $1/|T_r|$, not one. Define the realized graph return and its expectation by

$$F_r(\boldsymbol{\tau}) = (1-\lambda) \sum_v a_v R_v(\tau_v) + \lambda \sum_{e=(u,v)} b_e R_e(\tau_u, \tau_v), \tag{27}$$

$$J_r(\theta) = \mathbb{E}_\theta[F_r(\boldsymbol{\tau})], \qquad J(\theta) = \mathbb{E}_r[J_r(\theta)]. \tag{28}$$

The edge return contains only relational credit: endpoint rewards are counted in the node sum, not added again to every incident edge.

### B.2 Incident-edge credit and leave-one-out centering

For independent node samples $\tau_v^1, \ldots, \tau_v^G$, write $z_v^i = z_v(\tau_v^i)$ and $Q_{ij}^e = R_e(\tau_u^i, \tau_v^j)$ for $e = (u, v)$. Set

$$H_v^i = (1-\lambda) a_v R_v(\tau_v^i) + \lambda \sum_{e=(v,k)} b_e \frac{1}{G} \sum_{j=1}^G Q_{ij}^e + \lambda \sum_{e=(k,v)} b_e \frac{1}{G} \sum_{j=1}^G Q_{ji}^e, \tag{29}$$

$$B_v^i = \frac{1}{G-1} \sum_{k \neq i} H_v^k, \qquad \hat{A}_v^i = \mathrm{stopgrad}(H_v^i - B_v^i). \tag{30}$$

Both incoming and outgoing edges appear, with their original orientation in the reward arguments.

**Proposition 1** (Unbiased incident-edge credit and zero score-weighted baseline)**.** *Under the assumptions above,*

$$\mathbb{E}\left[\frac{1}{G} \sum_{v \in V_r} \sum_{i=1}^G z_v^i H_v^i\right] = \nabla_\theta J_r(\theta), \qquad \mathbb{E}[z_v^i B_v^i] = 0. \tag{31}$$

*Consequently, replacing $H_v^i$ by the detached advantage $\hat{A}_v^i$ preserves this expectation. Averaging over the fixed root distribution gives $\nabla_\theta J(\theta)$.*

*Proof.* For a node term, the score-function identity gives $\nabla_\theta \mathbb{E}[R_v] = \mathbb{E}[z_v R_v]$. For an edge $e = (u, v)$ with $u \neq v$, product sampling gives

$$\nabla_\theta \mathbb{E}[R_e(\tau_u, \tau_v)] = \mathbb{E}[(z_u + z_v) R_e(\tau_u, \tau_v)]. \tag{32}$$

Scores from every other node have zero expectation against that edge return by independence. Each pair $(\tau_u^i, \tau_v^j)$ has the same product distribution as $(\tau_u, \tau_v)$. Expanding the first estimator therefore

gives

$$\frac{1-\lambda}{G}\sum_{v,i} a_v z_v^i R_v(\tau_v^i) + \frac{\lambda}{G^2}\sum_{e=(u,v)} b_e \sum_{i,j}(z_u^i + z_v^j)Q_{ij}^e, \tag{33}$$

whose expectation is the sum of the derivatives in (28). Omitting either direction of incident credit would omit one of the two score terms in (32).

For the baseline, let $\mathcal{F}_{-vi}$ be the sigma-field generated by all sampled trajectories except $\tau_v^i$, conditional on the root and fixed worlds. If $k \neq i$, $H_v^k$ depends on $\tau_v^k$ and trajectories at neighboring nodes, but not on $\tau_v^i$: this is exactly where the absence of self-loops matters. Hence $B_v^i$ is $\mathcal{F}_{-vi}$-measurable and

$$\mathbb{E}[z_v^i B_v^i] = \mathbb{E}\big[B_v^i \mathbb{E}[z_v^i \mid \mathcal{F}_{-vi}]\big] = 0. \tag{34}$$

The other credits can share neighboring samples and be mutually correlated; their mutual independence is not needed. Independence of the excluded trajectory from the baseline is sufficient. Summing the zero baseline terms proves the result. □

The zero-expectation statement is about $z_v^i B_v^i$, not about $B_v^i$ itself. Rewards and advantages must be detached when differentiating a surrogate: differentiating through $H$ or $B$ would add terms absent from the proposition. A self-loop would reuse the node's own sample array inside its leave-one-out baseline and invalidate this argument, besides changing the product-policy interpretation of the edge.

**Relation to the implemented loss.** At the rollout policy $\theta_0$, let $r_{vit}(\theta) = \pi_\theta(x_{vit} \mid h_{vit}, w_v)/\pi_{\theta_0}(x_{vit} \mid h_{vit}, w_v)$ and compute $\hat{A}_v^i$ at $\theta_0$. The implemented clipped loss is

$$L(\theta) = -\frac{1}{G}\sum_{v,i}\sum_{t\in\mathcal{T}_{vi}} \min\{r_{vit}(\theta)\hat{A}_v^i, \mathrm{clip}(r_{vit}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_v^i\}. \tag{35}$$

At $\theta = \theta_0$, all ratios equal one and lie in the unclipped region, so $\mathbb{E}[-\nabla L(\theta_0)] = \nabla J_r(\theta_0)$. The loss is averaged over roots. There is no additional division by $|V_r|$, since $a_v$ and $b_e$ already normalize the objective. Only policy-generated tokens contribute; prompt, tool-observation, and padding tokens are masked. Token summation gives the trajectory score; trajectory-length averaging would define a different estimator. The implementation uses neither per-group standard-deviation normalization nor a default KL term. The result makes no unbiasedness claim for repeated off-policy updates, active clipping, or random group-standard-deviation scaling. The implementation checks in Table S13 are numerical consistency checks, not extensions of these assumptions.

### B.3 All-pair credit as a conditional mean of random matching

Let $\mathcal{D}$ denote all node trajectory arrays and their scores at a fixed policy. Independently of $\mathcal{D}$, draw a uniform random permutation $\sigma_e$ of $\{1, \ldots, G\}$ for each directed edge. The permutations are independent across edges in the matching construction. Define the *raw, uncentered* random-matching score estimator

$$\widehat{g}_{\mathrm{match}} = \frac{1-\lambda}{G}\sum_{v,i} a_v z_v^i R_v(\tau_v^i) + \frac{\lambda}{G}\sum_{e=(u,v)} b_e \sum_{i}(z_u^i + z_v^{\sigma_e(i)})Q_{i,\sigma_e(i)}^e. \tag{36}$$

This assigns every selected pair's return to both endpoints. Let $\widehat{g}_{\mathrm{all}}$ be (33).

**Proposition 2** (Conditional averaging for the raw score estimator)**.** *With the sampling and finite-second-moment assumptions above,*

$$\mathbb{E}[\widehat{g}_{\mathrm{match}} \mid \mathcal{D}] = \widehat{g}_{\mathrm{all}}, \qquad \mathrm{Cov}(\widehat{g}_{\mathrm{match}}) - \mathrm{Cov}(\widehat{g}_{\mathrm{all}}) \succeq 0. \tag{37}$$

*Both estimators have expectation $\nabla_\theta J_r(\theta)$.*

*Proof.* Conditioned on $\mathcal{D}$, scores and rewards are fixed. For every $e, i, j$, $\Pr(\sigma_e(i) = j \mid \mathcal{D}) = 1/G$, whence

$$\mathbb{E}\left[\frac{1}{G}\sum_i (z_u^i + z_v^{\sigma_e(i)})Q^e_{i,\sigma_e(i)} \;\middle|\; \mathcal{D}\right] = \frac{1}{G^2}\sum_{i,j}(z_u^i + z_v^j)Q^e_{ij}. \tag{38}$$

The node term is already measurable with respect to $\mathcal{D}$. Summing proves the conditional-mean identity, and Proposition 1 gives equality of expectations. The law of total covariance gives

$$\mathrm{Cov}(\widehat{g}_{\mathrm{match}}) = \mathrm{Cov}(\widehat{g}_{\mathrm{all}}) + \mathbb{E}[\mathrm{Cov}(\widehat{g}_{\mathrm{match}} \mid \mathcal{D})]. \tag{39}$$

The final term is positive semidefinite: for every deterministic vector $h$, its quadratic form is $\mathbb{E}[\mathrm{Var}(h^\top \widehat{g}_{\mathrm{match}} \mid \mathcal{D})] \geq 0$. □

The conditioning retains all dependence due to shared node samples across edges; no independence between the $G^2$ matrix entries is assumed. All-pair evaluation forms $|\mathcal{A}_r|G^2$ combinations from only $|V_r|G$ independently sampled trajectories. The proposition concerns the original score estimator without a baseline or random scale. We do not extend its covariance ordering to the implemented LOO estimator, normalized advantages, clipped losses, optimizer steps, or final policy performance. PTPO-Shared and PTPO are compared empirically for those quantities.

### B.4 Exact typed-pair decomposition and sampling weights

Let $d_v$ count all incoming and outgoing edges incident to $v$, and let $V_0 = \{v : d_v = 0\}$. For $e = (u, v)$ define

$$F_e(\tau_u, \tau_v) = (1-\lambda)\left(\frac{a_u}{d_u}R_u(\tau_u) + \frac{a_v}{d_v}R_v(\tau_v)\right) + \lambda b_e R_e(\tau_u, \tau_v). \tag{40}$$

No division by zero occurs: an endpoint of an edge has positive degree. The following identity holds pointwise:

$$F_r(\boldsymbol{\tau}) = \sum_{e\in\mathcal{A}_r} F_e(\tau_u, \tau_v) + (1-\lambda)\sum_{v\in V_0} a_v R_v(\tau_v). \tag{41}$$

Indeed, a nonisolated node occurs in exactly $d_v$ incident-edge terms, each with coefficient $(1-\lambda)a_v/d_v$; an isolated node occurs only in the final sum. Edge terms each occur once. Taking expectations establishes an exact decomposition of $J_r$.

If an edge $E$ is sampled with fixed probabilities $p_e > 0$ summing to one, then

$$\mathbb{E}_E\left[\frac{F_E}{p_E}\right] = \sum_e F_e. \tag{42}$$

The isolated-node sum must still be retained, either exactly or through a separate unbiased sampler. Equivalently, treat edges and isolated nodes as one collection of components, defining $F_v = (1-\lambda)a_v R_v$ for isolated nodes. Sampling a component $K$ with positive probability $q_K$ and returning $F_K/q_K$ estimates the entire return. For independent endpoint generation, an edge's score estimator is $(z_u + z_v)F_e/p_e$; its cross-node endpoint terms have expectation zero, so it also estimates the same objective gradient. Root sampling must preserve the declared root distribution, or use its own importance correction.

Thus a simple average over sampled edges without the importance factors generally reweights the objective. With the same rollout arrays, accumulating degree-distributed endpoint credit and all edge credit at each node before applying one node-level LOO operation reproduces (29) exactly. A single subsequent optimizer step then uses the same loss. Separate per-edge optimizer steps need not coincide. Independent typed-pair rollout generation preserves the objective but changes sample dependence and cost; the graph representation does not introduce a nondecomposable path or higher-order reward.

### B.5 What aggregate endpoint means do not identify

**Proposition 3** (Equal endpoint means, different aligned preservation)**.** *Even for a finite population with deterministic policies and valid syntax, the source and target population means of exact semantic endpoint accuracy do not determine instance-aligned Preserve-F1.*

*Proof.* Consider two equiprobable roots $r_1, r_2$, each with one Preserve edge $u \to v$. In each world let the gold semantic set be $\{a\}$; the target removes only irrelevant evidence. A stable evidence item $s$ supplies the sufficient proof $\{s\}$ for $a$ at both endpoints. Let $b \neq a$ be an unsupported atom. Correct predictions cite $s$; predictions of $b$ have invalid certificates, identically wherever they occur. Define two deterministic policies by the ordered pairs of semantic outputs

$$M: \quad (Y_{1u}, Y_{1v}) = (\{a\}, \{a\}), \qquad (Y_{2u}, Y_{2v}) = (\{b\}, \{b\}), \tag{43}$$

$$N: \quad (Y_{1u}, Y_{1v}) = (\{a\}, \{b\}), \qquad (Y_{2u}, Y_{2v}) = (\{b\}, \{a\}). \tag{44}$$

These policies can be realized on four distinct world prompts; deterministic trajectories satisfy product independence as point-mass distributions. For each policy, source endpoint accuracy is $1/2$ and target endpoint accuracy is $1/2$. In fact, at each endpoint the marginal distribution of predicted sets is also the same: half $\{a\}$ and half $\{b\}$. The same equality holds for the endpoint certificate-validity means under the certificate assignment above.

For $M$, the predicted intersection is $\{a\}$ on $r_1$ and $\{b\}$ on $r_2$, yielding Preserve-F1 values one and zero against $P_e^* = \{a\}$, hence mean $1/2$. For $N$, both predicted intersections are empty, yielding zero on both roots. All roots are eligible and each graph has one edge of one type, so the prescribed root/type aggregation yields the same means, $1/2$ and zero. Thus equal endpoint summaries admit different aligned responses. □

This is a non-identifiability statement about coarse population summaries, not independence from all endpoint information. The fully indexed endpoint outputs, gold sets, certificates, and graph determine the relational metrics. In particular, retaining root alignment distinguishes the two policies immediately. Likewise, exact complete semantic and certificate correctness at every endpoint already implies many of the corresponding relational constraints. Matched-endpoint experiments therefore address whether observed differences are summarized by the chosen aggregate endpoint measures; they do not establish statistical independence, hidden causal reasoning, or removal of all competence-related confounding.

## C Complete Experimental Tables

The following tables report complete protocol and diagnostic breakdowns (Tables S1–S24), detailed generalization results (Table S25), and mechanism controls (Table S26). All scores are on a 0–100 scale unless stated otherwise. Figure colors are chosen per figure family: the schematics (Figures 1 and 3) encode document and edge types, whereas the result figures encode methods and coordinates; the two palettes are independent. Unless a caption says otherwise, entries are means ± sample SD over five training seeds; the SD describes training reproducibility, not inferential uncertainty. A dash denotes an inapplicable aggregate, not a zero.

Table S1: Frozen training settings and selection thresholds. $\delta_{\text{match}}$ selects dev checkpoints; $\kappa$ assesses cross-method endpoint comparability; $\eta$ defines the non-inferiority margin. The MuSiQue matching triplet follows Table S28. These thresholds serve different purposes.

| Parameter | Frozen value |
|---|---|
| Stage-A lr / epochs | 1e-5 / 2 |
| RL lr | 1e-6 |
| clip $\epsilon$ | 0.20 |
| $\lambda$ | 0.50 |
| $G$ (rollouts per node) | 4 |
| Advantage | node-level LOO, no per-group std |
| KL coefficient | 0 |
| Temperature (train/eval) | 1.00 / 0.00 (greedy) |
| Max generation tokens | 1,024 |
| Graph batch | 8 roots per micro-batch, 16 micro-batches per update |
| RL steps | 2,000 |
| Checkpoint interval | 50 steps |
| QAMPARI tool budget | search/open/final, $\leq 6$ calls |
| Verifier timeout | 30 s |
| Random seeds | 17, 42, 73, 101, 2024 |
| $\delta_{\text{match}}$ (MuSiQue-Full) | $\pm 0.50$ / $\pm 1.00$ / $\pm 1.00$ |
| $\delta_{\text{match}}$ (FEVER, QAMPARI) | $\pm 0.50$ |
| $\kappa$ (matched-endpoint comparability) | 1.50 |
| $\eta$ (non-inferiority) | 1.00 |

Table S2: Audited derived graph counts, rather than original dataset sizes. Splits are assigned by root, with all associated worlds kept together. Worlds and edges per root are set by the intervention families applied to a root rather than sampled per split, so the per-root means are close across splits by construction; the counts are not independent draws.

| Dataset | Split | Roots | Worlds | Typed edges |
|---|---|---|---|---|
| MuSiQue-Full | train | 4,216 | 15,178 | 13,491 |
| | dev | 512 | 1,843 | 1,638 |
| | test | 528 | 1,901 | 1,690 |
| FEVER | train | 5,040 | 16,128 | 14,112 |
| | dev | 600 | 1,920 | 1,680 |
| | test | 640 | 2,048 | 1,792 |
| QAMPARI | train | 2,880 | 9,792 | 8,640 |
| | dev | 360 | 1,224 | 1,080 |
| | test | 380 | 1,292 | 1,140 |
| **Total** | train | 12,136 | 41,098 | 36,243 |
| **Total** | test | 1,548 | 5,241 | 4,622 |

Table S3: PTPO minus each baseline: mean differences with root-clustered paired bootstrap 95% intervals (10,000 resamples per seed; five-seed means); positive favors PTPO. Change/Preserve macros cover all datasets, Switch covers FEVER and QAMPARI, and Endpoint is a descriptive mixed-task macro. Only the Switch family is Holm-corrected; the other coordinates are exploratory. ‡ in panel (a) and † in panel (b) mark intervals containing zero. For Endpoint, only PTPO vs. GRPO-Marginal has a lower bound below the non-inferiority margin $-\eta$ ($\eta = 1.00$), so non-inferiority is not established for that contrast.

**(a) Change and Preserve**

| Comparison | Δ Change | Δ Preserve |
|---|---|---|
| PTPO vs SFT | $+13.12\,[+10.71, +15.53]$ | $+5.62\,[+4.33, +6.93]$ |
| PTPO vs GRPO-Marginal | $+6.00\,[+4.62, +7.38]$ | $+8.91\,[+6.86, +10.96]$ |
| PTPO vs CRPO-Evidence | $+0.79\,[-0.53, +2.11]^{\ddagger}$ | $+0.65\,[-0.75, +2.07]^{\ddagger}$ |
| PTPO vs Endpoint-LOO | $+5.41\,[+4.16, +6.66]$ | $+6.29\,[+4.84, +7.74]$ |
| PTPO vs Typed-Pair | $+1.94\,[+0.32, +3.56]$ | $+1.96\,[+0.33, +3.59]$ |
| PTPO vs PTPO-Shared | $+1.27\,[-0.18, +2.72]^{\ddagger}$ | $+1.12\,[-0.41, +2.65]^{\ddagger}$ |

**(b) Switch and Endpoint**

| Comparison | Δ Switch | Δ Endpoint |
|---|---|---|
| PTPO vs SFT | $+15.94\,[+12.20, +19.68]$ | $+4.84\,[+3.72, +5.96]$ |
| PTPO vs GRPO-Marginal | $+13.39\,[+9.90, +16.88]$ | $+0.14\,[-1.12, +1.40]^{\dagger}$ |
| PTPO vs CRPO-Evidence | $+4.28\,[+2.21, +6.35]$ | $+0.46\,[-0.28, +1.20]^{\dagger}$ |
| PTPO vs Endpoint-LOO | $+12.77\,[+9.62, +15.92]$ | $+0.53\,[-0.61, +1.67]^{\dagger}$ |
| PTPO vs Typed-Pair | $+5.48\,[+3.05, +7.91]$ | $+0.33\,[-0.62, +1.28]^{\dagger}$ |
| PTPO vs PTPO-Shared | $+3.69\,[+1.42, +5.96]$ | $+0.63\,[-0.19, +1.45]^{\dagger}$ |

Table S4: Frozen intervention counts and eligibility denominators. Panels (a-i) and (a-ii) separate test edge counts from metric denominators; panel (b) reports the training graph after the root-level holdouts of Table S18. Change requires nonempty $D^*$ and Preserve requires nonempty $P^*$; because a world's gold atom set may contain several atoms, Selective Set Drop can impose both obligations on the same edge and contributes to both coordinates, while Forced Switch contributes to Preserve and Switch. MuSiQue Switch is not applicable because no audited eligible alternative-proof family was obtained. The zero Forced Switch column in panel (b) follows from the full root-level holdout of that relation, not from a construction that never produced such edges.

**(a-i) Frozen test edge types**

| Dataset | Necessary drop | Irrelevant preserve | Selective set drop | Natural redundant | Forced Switch | Total |
|---|---|---|---|---|---|---|
| MuSiQue-Full | 640 | 610 | 176 | 264 | 0 | 1,690 |
| FEVER | 512 | 512 | 128 | 320 | 320 | 1,792 |
| QAMPARI | 285 | 285 | 230 | 175 | 165 | 1,140 |

**(a-ii) Frozen test metric denominators**

| Dataset | Change | Preserve | Natural adapt | Switch |
|---|---|---|---|---|
| MuSiQue-Full | 816 | 1,050 | 264 | — |
| FEVER | 640 | 1,280 | 320 | 320 |
| QAMPARI | 515 | 855 | 175 | 165 |

**(b) Frozen training split: edge types (no Forced Switch edges)**

| Dataset | Necessary drop | Irrelevant preserve | Selective set drop | Natural redundant | Forced Switch | Total |
|---|---|---|---|---|---|---|
| MuSiQue-Full | 5,126 | 4,857 | 1,417 | 2,091 | 0 | 13,491 |
| FEVER | 5,080 | 4,940 | 1,270 | 2,822 | 0 | 14,112 |
| QAMPARI | 2,402 | 2,299 | 2,298 | 1,641 | 0 | 8,640 |
| *Total* | 12,608 | 12,096 | 4,985 | 6,554 | 0 | 36,243 |

Table S5: Auxiliary T-Exact scores by dataset. Success requires complete semantic and certificate correctness at both endpoints and all applicable constraints. This conjunction is not used by itself to establish a distinct capability gap.

| Method | MuSiQue | FEVER | QAMPARI |
|---|---|---|---|
| SFT | 21.42 ±0.34 | 16.82 ±0.31 | 11.21 ±0.42 |
| GRPO-Marginal | 24.63 ±0.41 | 19.12 ±0.38 | 12.92 ±0.51 |
| CRPO-Evidence | 27.12 ±0.45 | 22.83 ±0.41 | 15.82 ±0.56 |
| **PTPO** | 27.93 ±0.42 | 24.12 ±0.39 | 16.72 ±0.54 |

Table S6: Error diagnostics and empty-set layers, macro-averaged across datasets; Base has no seed spread. UnexpectedAdd includes recovered correct content (not hallucination alone); NewUnsupported isolates unsupported additions. RecoveredPreserved: recovery; AnyRemoval and FalseDrop (erroneous removal of preserved gold) are both read on the $D^* = \emptyset$ layer. Panel (b) adds unsupported-content correction, the full-edge parse-failure rate, and the $D^* = \emptyset$-layer union of format failure with FalseDrop; Preserve is not applicable when $P^* = \emptyset$. A union involving the layer-restricted FalseDrop is itself a layer quantity and is not an all-edge statistic: combining it with the full-edge parse-failure rate would exceed the all-edge bound $P(F) + P(D \mid D^* = \emptyset)\,|D^* = \emptyset|/|E|$. The FalseDrop row repeats the same $D^* = \emptyset$-layer quantity as panel (a-ii).

**(a-i) Additions and recovery**

| Method | Unexpected add | New unsupported | Recovered preserved |
|---|---|---|---|
| Base | 19.32 | 16.41 | 2.13 |
| SFT | 12.43 ±0.31 | 8.92 ±0.26 | 3.51 ±0.14 |
| GRPO-Marginal | 17.63 ±0.44 | 13.21 ±0.38 | 4.42 ±0.19 |
| Endpoint-LOO | 15.12 ±0.39 | 11.43 ±0.33 | 4.13 ±0.17 |
| CRPO-Evidence | 11.21 ±0.29 | 7.62 ±0.24 | 3.62 ±0.15 |
| **PTPO** | 10.42 ±0.27 | 7.13 ±0.22 | 3.32 ±0.13 |

**(a-ii) Removal and parsing**

| Method | FalseDrop | AnyRemoval | Format-fail |
|---|---|---|---|
| Base | 12.63 | 37.42 | 14.91 |
| SFT | 9.72 ±0.29 | 41.23 ±0.42 | 2.81 ±0.11 |
| GRPO-Marginal | 14.83 ±0.47 | 58.61 ±0.66 | 2.12 ±0.10 |
| Endpoint-LOO | 12.92 ±0.41 | 54.12 ±0.58 | 2.21 ±0.10 |
| CRPO-Evidence | 7.91 ±0.26 | 44.83 ±0.48 | 1.92 ±0.09 |
| **PTPO** | 7.21 ±0.24 | 43.62 ±0.45 | 1.81 ±0.08 |

**(b) Empty-set layers and parse failures (macro over three datasets)**

| Layer | Metric | SFT | GRPO-Marginal | PTPO |
|---|---|---|---|---|
| $D^* = \emptyset$ edges | AnyRemoval | 41.23 ±0.42 | 58.61 ±0.66 | 43.62 ±0.45 |
| | FalseDrop | 9.72 ±0.29 | 14.83 ±0.47 | 7.21 ±0.24 |
| | CorrectedUnsupported | 18.42 ±0.38 | 21.32 ±0.44 | 22.91 ±0.48 |
| | Format-fail ∪ FalseDrop | 11.92 ±0.35 | 16.42 ±0.52 | 8.82 ±0.26 |
| $P^* = \emptyset$ edges | Preserve | — | — | — |
| All edges | Format-fail | 2.81 ±0.11 | 2.12 ±0.10 | 1.81 ±0.08 |

Table S7: No-op and sham-edit controls, macro-averaged across datasets. No-op repeats the same world, prompt and checkpoint; sham edits change text without changing support; real removals are the reference. The residual nonzero change of a no-op is an upper bound on nuisance variation under greedy decoding, and these controls alone do not rule out all editing or execution confounds.

| Condition | Semantic change (%) | Certificate invalid (%) |
|---|---|---|
| No-op (repeated world) | 3.12 ±0.21 | 1.21 ±0.13 |
| Sham edit (matched edit volume) | 5.42 ±0.29 | 2.12 ±0.18 |
| Irrelevant removal (real) | 6.82 ±0.34 | 3.41 ±0.24 |
| Necessary removal (real) | 74.86 ±1.24 | 69.42 ±1.36 |

Table S8: Training cost and generation-budget controls. (a) Costs are per seed; RL rows use the reported full-step budget and Stage-A is listed separately. Rollouts count every world-level policy sampling over the full budget: the eight-root batch is a micro-batch, sixteen micro-batches are accumulated per update, and with $G = 4$ the total is $2{,}000 \times 128 \times \bar{n}_w \times 4 \approx 3.5 \times 10^6$ world rollouts, so the count is not the number of updates times the micro-batch size. Verifier calls cover edge-level typed reward evaluation; wall time is GPU hours divided by the allocated devices and is not an independent measurement. (b) Typed-Pair with independent sampling at full and matched generation budgets, with PTPO as reference. Equal token budgets do not imply equal update counts. Outcomes follow Table S26.

**(a) Training cost on the main backbone, per seed, 2,000 RL steps**

| Method | Rollouts (M) | Generated tokens (B) | Verifier calls (M) | Wall time (h) | GPU hours |
|---|---|---|---|---|---|
| Stage-A SFT | 0.31 ±0.00 | 0.42 ±0.01 | 0.00 | 17.08 ±0.36 | 68.31 ±1.42 |
| GRPO-Marginal | 3.55 ±0.02 | 4.80 ±0.04 | 0.00 | 103.12 ±1.90 | 412.47 ±7.61 |
| Endpoint-LOO | 3.55 ±0.02 | 4.80 ±0.04 | 0.00 | 103.78 ±1.84 | 415.13 ±7.34 |
| CRPO-Evidence | 3.56 ±0.02 | 4.82 ±0.04 | 3.61 ±0.05 | 121.56 ±2.26 | 486.24 ±9.03 |
| Typed-Pair | 4.39 ±0.03 | 5.94 ±0.05 | 3.61 ±0.05 | 143.55 ±2.72 | 574.18 ±10.87 |
| PTPO-Shared | 3.56 ±0.02 | 4.82 ±0.04 | 3.61 ±0.05 | 122.36 ±2.19 | 489.42 ±8.76 |
| **PTPO** | 3.56 ±0.02 | 4.82 ±0.04 | 3.61 ±0.05 | 122.84 ±2.31 | 491.36 ±9.24 |

**(b) Matched generation budget: Typed-Pair at 4.82B tokens versus PTPO**

| Configuration | Budget | MuSiQue Change | MuSiQue Preserve | FEVER Switch | QAMPARI Switch |
|---|---|---|---|---|---|
| Typed-Pair | full budget (5.94B tokens) | 55.42 ±0.53 | 61.62 ±0.39 | 30.31 ±0.59 | 25.58 ±0.94 |
| Typed-Pair | matched budget (4.82B tokens) | 54.51 ±0.55 | 60.82 ±0.41 | 29.13 ±0.61 | 24.36 ±0.97 |
| **PTPO** | 4.82B tokens | 57.24 ±0.39 | 63.52 ±0.41 | 35.94 ±0.50 | 30.91 ±0.82 |

Table S9: Natural-redundancy adaptation and conditional switching. (a) SupportAdaptAcc over all eligible natural-redundancy edges with nonempty frozen target sets (MuSiQue 264, FEVER 320, QAMPARI 175); Macro averages datasets. (b) SwitchCond conditions on each method's triggered target units and reports trigger coverage; these method-dependent subsets support diagnosis, not a common-denominator ranking. MuSiQue has natural-redundancy edges but no eligible Forced Switch family, so its Switch column is not applicable (Table S4).

**(a) Natural-redundancy adaptation (SupportAdaptAcc)**

| Method | MuSiQue | FEVER | QAMPARI | Macro |
|---|---|---|---|---|
| SFT | 51.92 ±0.46 | 58.41 ±0.42 | 46.32 ±0.61 | 52.22 ±0.44 |
| GRPO-Marginal | 49.21 ±0.52 | 55.13 ±0.48 | 43.72 ±0.72 | 49.35 ±0.51 |
| CRPO-Evidence | 65.12 ±0.55 | 70.23 ±0.50 | 58.92 ±0.78 | 64.76 ±0.54 |
| **PTPO** | 68.62 ±0.49 | 72.91 ±0.46 | 61.72 ±0.74 | 67.75 ±0.50 |

**(b) Conditional switching diagnosis (triggered units only)**

| | FEVER | | QAMPARI | |
|---|---|---|---|---|
| Method | SwitchCond | Coverage | SwitchCond | Coverage |
| SFT | 34.82 ±0.94 | 31.21 ±0.78 | 28.41 ±1.12 | 26.92 ±0.85 |
| GRPO-Marginal | 37.62 ±1.02 | 34.52 ±0.84 | 30.13 ±1.21 | 29.41 ±0.92 |
| CRPO-Evidence | 48.92 ±1.15 | 52.13 ±0.96 | 42.32 ±1.38 | 47.82 ±1.05 |
| **PTPO** | 55.12 ±1.08 | 58.42 ±0.91 | 48.62 ±1.31 | 52.13 ±0.99 |

Table S10: Per-seed macro coordinates and reported summaries. Change/Preserve cover all datasets; Switch covers FEVER and QAMPARI. Endpoint mixes task-specific semantic scales and is descriptive only. SD is the sample SD, not a confidence interval; statistical comparisons are in Table S3.

| Method | Coordinate | s17 | s42 | s73 | s101 | s2024 | Mean | SD |
|---|---|---|---|---|---|---|---|---|
| SFT | Change | 44.92 | 44.45 | 44.32 | 44.64 | 44.84 | 44.63 | 0.25 |
| | Preserve | 61.24 | 60.79 | 60.65 | 60.95 | 61.14 | 60.95 | 0.24 |
| | Switch | 17.83 | 17.28 | 16.92 | 17.61 | 17.77 | 17.48 | 0.38 |
| | Endpoint | 47.08 | 46.68 | 46.57 | 46.83 | 47.00 | 46.83 | 0.21 |
| GRPO-Marginal | Change | 52.01 | 51.64 | 51.60 | 51.91 | 51.61 | 51.75 | 0.19 |
| | Preserve | 57.95 | 57.57 | 57.50 | 57.80 | 57.51 | 57.67 | 0.20 |
| | Switch | 20.49 | 19.78 | 20.18 | 20.07 | 19.66 | 20.04 | 0.33 |
| | Endpoint | 51.73 | 51.46 | 51.39 | 51.63 | 51.42 | 51.53 | 0.15 |
| Endpoint-LOO | Change | 52.68 | 52.02 | 52.71 | 52.17 | 52.12 | 52.34 | 0.33 |
| | Preserve | 60.58 | 60.01 | 60.62 | 60.15 | 60.10 | 60.29 | 0.29 |
| | Switch | 21.03 | 19.98 | 21.19 | 20.52 | 20.55 | 20.65 | 0.48 |
| | Endpoint | 51.36 | 50.95 | 51.40 | 51.03 | 50.98 | 51.14 | 0.22 |
| CRPO-Evidence | Change | 56.92 | 57.28 | 56.93 | 56.63 | 57.04 | 56.96 | 0.23 |
| | Preserve | 65.93 | 66.17 | 65.84 | 65.69 | 65.98 | 65.92 | 0.18 |
| | Switch | 28.81 | 29.85 | 28.64 | 29.03 | 29.40 | 29.15 | 0.49 |
| | Endpoint | 51.17 | 51.44 | 51.14 | 51.01 | 51.28 | 51.21 | 0.16 |
| Typed-Pair | Change | 55.86 | 55.56 | 56.39 | 56.27 | 54.98 | 55.81 | 0.57 |
| | Preserve | 64.66 | 64.43 | 65.05 | 64.96 | 64.01 | 64.62 | 0.42 |
| | Switch | 28.07 | 27.69 | 28.79 | 28.33 | 26.83 | 27.94 | 0.74 |
| | Endpoint | 51.38 | 51.17 | 51.74 | 51.65 | 50.78 | 51.34 | 0.39 |
| PTPO-Shared | Change | 56.87 | 56.06 | 56.87 | 56.30 | 56.32 | 56.48 | 0.37 |
| | Preserve | 65.78 | 65.12 | 65.78 | 65.30 | 65.34 | 65.46 | 0.30 |
| | Switch | 30.52 | 29.31 | 30.21 | 29.47 | 29.14 | 29.73 | 0.60 |
| | Endpoint | 51.32 | 50.81 | 51.28 | 50.91 | 50.89 | 51.04 | 0.24 |
| **PTPO** | Change | 57.92 | 57.85 | 57.96 | 57.33 | 57.71 | 57.75 | 0.26 |
| | Preserve | 66.74 | 66.69 | 66.74 | 66.18 | 66.53 | 66.58 | 0.24 |
| | Switch | 33.54 | 33.28 | 33.92 | 32.88 | 33.49 | 33.42 | 0.38 |
| | Endpoint | 51.81 | 51.76 | 51.78 | 51.40 | 51.59 | 51.67 | 0.17 |

Table S11: Behavior by affected-answer count, proof hops, alternative-proof families and edit volume; five-seed means. Counts refer to Change-eligible edges except in the Switch stratum; Preserve entries use the subset that also satisfies Preserve eligibility. Affected-answer strata count atoms of $D_e^*$ (cf. Table S4). Panels (a), (b), and (d) pool the eligible edges directly, whereas the Change and Preserve coordinates of Table 1 average edges within a root and primary type before averaging roots and types; those strata therefore do not reweight to the reported coordinates. Panel (c) uses the same edge-pooled aggregation as SwitchAcc and therefore reweights the two tasks by eligible-edge count (320 vs. 165); its pooled totals are 18.19 for SFT and 34.23 for PTPO, which differ by construction from the equal-task Switch macro of Table 1 (17.48 and 33.42). Panel (c) counts alternative-proof families per edge, whereas Table S25(b) holds out the per-target predicate of at least three alternative proofs at the root level; the two edge sets differ and are not nested.

**(a) By number of affected answers (Change denominator 1,971)**

| Affected answers | Edges | SFT Change | PTPO Change | SFT Preserve | PTPO Preserve |
|---|---|---|---|---|---|
| 1 | 842 | 45.31 | 57.82 | 60.42 | 66.13 |
| 2–3 | 613 | 41.27 | 54.16 | 58.71 | 65.24 |
| 4–6 | 331 | 36.42 | 48.93 | 55.18 | 62.47 |
| $\geq 7$ | 185 | 30.84 | 42.61 | 51.32 | 58.92 |

**(b) By proof hop (Change denominator 1,971)**

| Hop | Edges | SFT Change | PTPO Change | SFT Preserve | PTPO Preserve |
|---|---|---|---|---|---|
| 2 | 604 | 46.83 | 59.14 | 61.72 | 67.31 |
| 3 | 826 | 43.92 | 56.48 | 59.84 | 65.93 |
| $\geq 4$ | 541 | 36.14 | 48.27 | 53.16 | 60.12 |

**(c) By number of alternative-proof families (Switch denominator 485)**

| Alternative proofs | Edges | SFT Switch | PTPO Switch |
|---|---|---|---|
| 1 | 186 | 23.66 | 40.65 |
| 2 | 172 | 17.56 | 33.49 |
| $\geq 3$ | 127 | 11.02 | 25.83 |

**(d) By intervention edit volume (Change denominator 1,971)**

| Edit volume | Edges | SFT Change | PTPO Change | SFT Preserve | PTPO Preserve |
|---|---|---|---|---|---|
| Small ($< 10\%$) | 721 | 47.16 | 59.83 | 62.14 | 67.82 |
| Medium (10–30%) | 836 | 43.28 | 55.72 | 59.31 | 65.14 |
| Large ($> 30\%$) | 414 | 35.91 | 47.16 | 52.84 | 59.73 |

Table S12: Errors conditional on a correct source answer, macro-averaged across datasets; five-seed means. Coverage is reported separately; disjoint error categories sum to the conditional error rate. Similar coverage does not guarantee identical subsets, so these diagnostics do not replace full-denominator comparisons.

| Method | Source-correct coverage | Conditional error | Over-retraction | Spurious retention | Certificate repair failure |
|---|---|---|---|---|---|
| SFT | 62.14 | 48.31 | 18.42 | 21.63 | 8.26 |
| GRPO-Marginal | 71.83 | 45.72 | 24.91 | 12.48 | 8.33 |
| Endpoint-LOO | 70.94 | 44.16 | 23.71 | 12.63 | 7.82 |
| CRPO-Evidence | 70.42 | 38.14 | 15.23 | 14.82 | 8.09 |
| **PTPO** | 71.26 | 35.87 | 13.41 | 14.62 | 7.84 |

Table S13: Implementation consistency checks reported on frozen graphs and checkpoints. Monte Carlo checks use $10^5$ draws. These are numerical consistency checks rather than new theoretical results. The edge-weight row sums $b_e = 1/(|T_r|n_{\rho_e})$ of Section 5 over every edge of a root graph, matching $\sum_e b_e = 1$; the per-type partial sum is $1/|T_r|$ and is not separately checked here. The LOO row checks the score-weighted expectation $\mathbb{E}[z_v^i B_v^i] = 0$ of Proposition 1 (Appendix B), not the (generally nonzero) expectation of the baseline $B_v^i$ itself.

| Check | Expected | Measured | Verdict |
|---|---|---|---|
| Node-weight sum $\Sigma a_v$ per root graph | 1.000000 | $1.000000 \pm 3\text{e-}7$ | pass |
| Edge-weight sum $\Sigma b_e$, root graph total | 1.000000 | $1.000000 \pm 4\text{e-}7$ | pass |
| Self-loop edges | 0 | 0 | pass |
| In-edge + out-edge credit coverage | 100% | 100.00% | pass |
| Graph vs same-batch weighted-pair, max per-parameter gap | $< 10^{-6}$ | 3.2e-7 | pass |
| $\lambda = 0$ credit scale, relative error | $< 10^{-9}$ | 4.1e-10 | pass |
| Loss gap after node-order shuffling | 0 | 0.0 | pass |
| Loss gap after edge-sum-order shuffling | 0 | 0.0 | pass |
| Advantage gap after within-group shuffling | 0 | 0.0 | pass |
| Random-matching estimator, conditional mean (MC $10^5$) | unbiased | $+0.0008 \pm 0.0021$ | pass |
| Score-weighted LOO baseline $\mathbb{E}[z \cdot B]$, frozen policy | 0 | $+0.0003 \pm 0.0018$ | pass |

Table S14: QAMPARI retrieval-interface results. Tool failures (budget exhaustion, timeout, or malformed action) remain in the denominator. Every row uses the QAMPARI retrieval interface (`search`/`open`/`final`, at most six calls); the static direct-context results for the same task are in Tables 1 and S25(c). The static→agent row reuses the MuSiQue-trained policy, and the in-domain row is trained and evaluated under the agent interface. Tool-failure rates differ across settings and are reported so that residual losses are not attributed to retrieval alone.

| Setting | Endpoint | Change | Preserve | Switch | Tool failure (%) |
|---|---|---|---|---|---|
| SFT (agent interface) | 24.81 ±0.52 | 35.42 ±0.61 | 56.28 ±0.55 | 13.82 ±0.94 | 3.12 ±0.14 |
| **PTPO (static → agent)** | 25.93 ±0.49 | 38.14 ±0.58 | 57.61 ±0.52 | 17.45 ±0.88 | 2.64 ±0.11 |
| **PTPO (agent in-domain)** | 27.64 ±0.46 | 45.23 ±0.54 | 61.42 ±0.48 | 26.18 ±0.81 | 1.87 ±0.09 |

Table S15: Per-seed PTPO-minus-baseline differences on macro Change and Switch. Change averages all datasets; Switch averages FEVER and QAMPARI. Sign consistency across seeds describes training variation and does not replace root-clustered inference in Table S3.

**(a) Change differences**

| Seed | vs SFT | vs GRPO-Marginal | vs CRPO-Evidence | vs Typed-Pair |
|---|---|---|---|---|
| s17 | +13.00 | +5.91 | +1.00 | +2.06 |
| s42 | +13.40 | +6.21 | +0.57 | +2.29 |
| s73 | +13.64 | +6.36 | +1.03 | +1.57 |
| s101 | +12.69 | +5.42 | +0.70 | +1.06 |
| s2024 | +12.87 | +6.10 | +0.67 | +2.73 |

**(b) Switch differences**

| Seed | vs SFT | vs GRPO-Marginal | vs CRPO-Evidence |
|---|---|---|---|
| s17 | +15.71 | +13.05 | +4.73 |
| s42 | +16.00 | +13.50 | +3.43 |
| s73 | +17.00 | +13.74 | +5.28 |
| s101 | +15.27 | +12.81 | +3.85 |
| s2024 | +15.72 | +13.83 | +4.09 |

Table S16: Reported root-clustered bootstrap characteristics. Effective sample sizes summarize dependence and do not replace the specified aggregation estimator. The Switch dataset pool is FEVER and QAMPARI; its Roots entry is the sum of their test roots, the pool over which the switch contrasts are resampled. Kernel evaluations are not independent experimental samples.

| Coordinate | Macro denom. | Roots | Design effect | Effective $n$ | Bootstrap SE |
|---|---|---|---|---|---|
| Change | 1,971 | 1,548 | 1.58 | 1,247 | 1.35 |
| Preserve | 3,185 | 1,548 | 1.66 | 1,918 | 1.45 |
| Switch | 485 | 1,020 | 1.52 | 319 | 2.10 |
| Endpoint | — | 1,548 | 1.41 | 1,098 | 0.75 |

Table S17: Reported test statistics. Switch contrasts form the Holm-corrected primary family; Change, Preserve, and Endpoint values are uncorrected exploratory tests. All $p$-values are two-sided Wald values computed from the paired root-clustered bootstrap standard errors that set the interval widths in Table S3; they are not read off the marginal per-coordinate standard errors of Table S16. Because no cross-task Endpoint aggregate is formed, the Endpoint column evaluates each task separately and reports the least favorable (largest) of the three. Endpoint difference-test $p$-values do not establish non-inferiority; interpretation requires the stated margin and the interval bounds of Table S3, where only PTPO vs GRPO-Marginal has an Endpoint lower bound below the non-inferiority margin.

| Comparison | $p$ (Change) | $p$ (Preserve) | $p$ (Switch) raw $\rightarrow$ Holm | $p$ (Endpoint) |
|---|---|---|---|---|
| PTPO vs SFT | $< 10^{-16}$ | $< 10^{-16}$ | $< 10^{-16} \rightarrow < 10^{-16}$ | $< 10^{-16}$ |
| PTPO vs GRPO-Marginal | $< 10^{-16}$ | $< 10^{-16}$ | 5.60e-14 $\rightarrow$ 2.24e-13 | 0.828 |
| PTPO vs CRPO-Evidence | 0.241 | 0.366 | 5.12e-05 $\rightarrow$ 1.02e-04 | 0.223 |
| PTPO vs Endpoint-LOO | $< 10^{-16}$ | $< 10^{-16}$ | 1.78e-15 $\rightarrow$ 8.88e-15 | 0.362 |
| PTPO vs Typed-Pair | 0.019 | 0.018 | 9.90e-06 $\rightarrow$ 2.97e-05 | 0.496 |
| PTPO vs PTPO-Shared | 0.086 | 0.151 | 1.43e-03 $\rightarrow$ 1.43e-03 | 0.132 |

Table S18: Exposure exclusions. Held-out objects are removed at the root level: all worlds and typed edges of a held-out root leave the training graph. Shares use the 36,243 training edges of Table S4(b). Forced Switch is held out in full (hence no Switch edges in that panel) and is evaluated in Table S25(a). Stage-A single-world supervision is shared by all methods within a seed, so holdouts act on the RL graph only.

| Held-out object | Roots | Nodes | Edges | Held-out edge share |
|---|---|---|---|---|
| Forced Switch relation (Table S25(a)) | 412 | 1,483 | 1,286 | 3.55% |
| Joint multi-sentence proof (Table S25(b)) | 268 | 964 | 842 | 2.32% |
| Long hop $\geq 4$ (Table S25(b)) | 341 | 1,228 | 1,064 | 2.94% |
| $\geq 3$ alternative proofs (Table S25(b)) | 156 | 561 | 487 | 1.34% |

Table S19: Implied paired correlations, derived rather than directly measured. The reported construction combines interval half-widths from Table S3 with single-method standard errors from Table S16, assuming a common marginal standard error within each contrast. These derived values do not establish matching quality or enter significance decisions.

| Comparison | $\rho$ (Change) | $\rho$ (Preserve) | $\rho$ (Switch) |
|---|---|---|---|
| PTPO vs SFT | 0.59 | 0.90 | 0.59 |
| PTPO vs GRPO-Marginal | 0.86 | 0.74 | 0.64 |
| PTPO vs CRPO-Evidence | 0.88 | 0.88 | 0.87 |
| PTPO vs Endpoint-LOO | 0.89 | 0.87 | 0.71 |
| PTPO vs Typed-Pair | 0.81 | 0.84 | 0.83 |
| PTPO vs PTPO-Shared | 0.85 | 0.86 | 0.85 |

Table S20: Task-native scores and supporting diagnostics, evaluated on the same frozen domain that supplies the Endpoint column of Table 1: panel (a) answer F1, panel (b) label recall, and panel (c) answer F1 therefore reproduce those Endpoint values, and the remaining columns carry companion diagnostics that the main table does not report. MuSiQue: answer/support and group-sufficiency; FEVER: label recall vs. evidence scores; QAMPARI precision/recall measure answer correctness, not certificate validity. Question-level averaging need not preserve the harmonic identity. Bold marks the focal PTPO row, not the maximum.

**(a) MuSiQue-Full: official-subset answer and support metrics**

| Method | Answer F1 | Support F1 | Evidence valid (%) | Group answer sufficiency F1 | Group support sufficiency F1 |
|---|---|---|---|---|---|
| SFT | 35.14 ±0.23 | 48.62 ±0.31 | 48.31 ±0.34 | 33.81 ±0.29 | 45.27 ±0.33 |
| GRPO-Marginal | 40.82 ±0.33 | 51.34 ±0.38 | 51.27 ±0.41 | 39.42 ±0.36 | 48.15 ±0.40 |
| Endpoint-LOO | 40.41 ±0.27 | 51.02 ±0.35 | 50.84 ±0.38 | 39.05 ±0.33 | 47.86 ±0.37 |
| CRPO-Evidence | 40.13 ±0.31 | 52.87 ±0.42 | 54.62 ±0.45 | 38.74 ±0.39 | 49.63 ±0.44 |
| Typed-Pair | 40.58 ±0.35 | 52.41 ±0.44 | 53.91 ±0.47 | 39.18 ±0.41 | 49.12 ±0.46 |
| PTPO-Shared | 40.29 ±0.29 | 52.63 ±0.40 | 54.38 ±0.43 | 38.91 ±0.37 | 49.38 ±0.42 |
| **PTPO** | 40.71 ±0.31 | 53.42 ±0.39 | 55.17 ±0.42 | 39.33 ±0.35 | 50.16 ±0.41 |

**(b) FEVER: label recall, FEVER-style score, evidence validity**

| Method | Label recall | FEVER score | Evidence valid (%) |
|---|---|---|---|
| SFT | 78.94 ±0.22 | 61.42 ±0.34 | 66.83 ±0.36 |
| GRPO-Marginal | 83.62 ±0.25 | 64.81 ±0.39 | 69.24 ±0.41 |
| Endpoint-LOO | 83.21 ±0.30 | 64.53 ±0.37 | 68.92 ±0.39 |
| CRPO-Evidence | 83.87 ±0.34 | 68.14 ±0.43 | 73.46 ±0.45 |
| Typed-Pair | 83.41 ±0.36 | 67.52 ±0.45 | 72.81 ±0.47 |
| PTPO-Shared | 83.13 ±0.34 | 67.93 ±0.41 | 73.12 ±0.43 |
| **PTPO** | 84.02 ±0.30 | 68.71 ±0.42 | 74.05 ±0.44 |

**(c) QAMPARI: answer-set precision, recall and output shape**

| Method | Answer F1 | Precision | Recall | Empty answer (%) | Mean predicted count |
|---|---|---|---|---|---|
| SFT | 26.42 ±0.31 | 31.24 ±0.36 | 27.83 ±0.34 | 8.42 ±0.29 | 4.62 ±0.08 |
| GRPO-Marginal | 30.13 ±0.44 | 33.87 ±0.48 | 32.41 ±0.46 | 5.21 ±0.31 | 5.87 ±0.11 |
| Endpoint-LOO | 29.81 ±0.38 | 33.52 ±0.43 | 32.08 ±0.41 | 5.46 ±0.28 | 5.74 ±0.10 |
| CRPO-Evidence | 29.62 ±0.54 | 34.91 ±0.56 | 31.72 ±0.53 | 4.13 ±0.33 | 5.18 ±0.12 |
| Typed-Pair | 30.04 ±0.53 | 34.52 ±0.55 | 32.15 ±0.52 | 4.38 ±0.32 | 5.31 ±0.12 |
| PTPO-Shared | 29.71 ±0.43 | 34.74 ±0.46 | 31.86 ±0.44 | 4.25 ±0.30 | 5.24 ±0.11 |
| **PTPO** | 30.28 ±0.41 | 35.63 ±0.44 | 32.47 ±0.42 | 3.81 ±0.29 | 5.09 ±0.10 |

Table S21: Precision, recall, and F1 of Change and Preserve, with the same root/type aggregation and dataset macro as Table 1. F1 is averaged independently at the edge level, rather than recomputed from macro precision and recall. PTPO holds the highest precision, recall, and F1 on both coordinates.

**(a) Change**

| Method | Precision | Recall | F-score |
|---|---|---|---|
| SFT | 51.80 ±0.31 | 42.15 ±0.28 | 44.63 ±0.25 |
| GRPO-Marginal | 57.42 ±0.24 | 50.16 ±0.21 | 51.75 ±0.19 |
| Endpoint-LOO | 58.94 ±0.38 | 48.72 ±0.34 | 52.34 ±0.33 |
| CRPO-Evidence | 63.85 ±0.28 | 53.41 ±0.25 | 56.96 ±0.23 |
| Typed-Pair | 61.42 ±0.62 | 54.28 ±0.58 | 55.81 ±0.57 |
| PTPO-Shared | 63.28 ±0.41 | 52.94 ±0.38 | 56.48 ±0.37 |
| **PTPO** | 65.13 ±0.29 | 54.62 ±0.27 | 57.75 ±0.26 |

**(b) Preserve**

| Method | Precision | Recall | F-score |
|---|---|---|---|
| SFT | 57.42 ±0.28 | 65.13 ±0.26 | 60.95 ±0.24 |
| GRPO-Marginal | 54.18 ±0.23 | 61.72 ±0.21 | 57.67 ±0.20 |
| Endpoint-LOO | 56.83 ±0.33 | 64.41 ±0.30 | 60.29 ±0.29 |
| CRPO-Evidence | 62.20 ±0.21 | 70.15 ±0.19 | 65.92 ±0.18 |
| Typed-Pair | 60.85 ±0.48 | 69.10 ±0.44 | 64.62 ±0.42 |
| PTPO-Shared | 61.60 ±0.35 | 69.90 ±0.32 | 65.46 ±0.30 |
| **PTPO** | 62.40 ±0.28 | 71.50 ±0.26 | 66.58 ±0.24 |

Table S22: Unit-level switching rates and mean targets per edge. A target is one semantic atom of the frozen set $K_e$, so a world with several gold atoms contributes several targets on a single edge (hence means above one); target counts are fixed by construction and carry no seed spread. Both columns share the same frozen Switch-eligible edge set and the same full-edge denominator. Edge-level SwitchAcc scores an edge 1 only if every target succeeds and neither endpoint carries an extra incorrect semantic prediction, whereas the unit-level rate reported here averages each edge's per-target success fraction with equal edge weight and applies no such output-level gate, $U_{\text{edge}} = \frac{1}{|\mathcal{E}_{\text{sw}}|} \sum_e \frac{1}{|K_e|} \sum_{a \in K_e} h_e(a)$ (Eq. (13)); the two share eligibility but differ in aggregation order and in this gate. Because an all-succeed indicator never exceeds that edge's own per-target average, edge-level SwitchAcc is necessarily below the unit-level rate.

| Method | FEVER unit-level | QAMPARI unit-level | Targets/edge FEVER / QAMPARI |
|---|---|---|---|
| SFT | 28.42 ±0.44 | 24.17 ±0.58 | 2.14 / 2.86 |
| GRPO-Marginal | 31.86 ±0.62 | 27.53 ±0.84 | 2.14 / 2.86 |
| Endpoint-LOO | 32.41 ±0.59 | 28.06 ±0.81 | 2.14 / 2.86 |
| CRPO-Evidence | 42.73 ±0.71 | 37.84 ±0.94 | 2.14 / 2.86 |
| Typed-Pair | 41.28 ±0.64 | 36.42 ±0.97 | 2.14 / 2.86 |
| PTPO-Shared | 43.16 ±0.60 | 38.21 ±1.03 | 2.14 / 2.86 |
| **PTPO** | 47.05 ±0.53 | 41.63 ±0.88 | 2.14 / 2.86 |

Table S23: Development checkpoint selection underlying Table S28. Selection is independent per seed on development data within the predeclared common endpoint range and uses no test response score; ties prefer fewer training tokens and then the fixed checkpoint identifier. The MuSiQue-Full triplet is sufficient-node answer F1, sufficient-node answerability, and insufficient-node answerability; the other datasets match their primary semantic endpoint. Rows list the methods that enter the matched comparison: four on MuSiQue-Full, three on FEVER and QAMPARI.

| Dataset | Method | Dev step (range) | Dev matched metric |
|---|---|---|---|
| MuSiQue-Full | GRPO-Marginal | 1,600 (1,550–1,650) | 39.81 / 82.43 / 74.12 |
| | Endpoint-LOO | 1,650 (1,600–1,700) | 39.62 / 82.14 / 74.61 |
| | CRPO-Evidence | 1,500 (1,450–1,550) | 39.66 / 82.05 / 74.71 |
| | **PTPO** | 1,450 (1,400–1,500) | 39.91 / 82.61 / 74.42 |
| FEVER | GRPO-Marginal | 1,700 (1,650–1,750) | 82.91 |
| | CRPO-Evidence | 1,550 (1,500–1,600) | 83.18 |
| | **PTPO** | 1,500 (1,450–1,550) | 83.12 |
| QAMPARI | GRPO-Marginal | 1,650 (1,600–1,700) | 29.82 |
| | CRPO-Evidence | 1,550 (1,500–1,600) | 29.51 |
| | **PTPO** | 1,500 (1,450–1,550) | 30.12 |

Table S24: Greedy and temperature-1.00 ($T$=1.00) evaluations of the same checkpoints and frozen graphs; five-seed means. Change/Preserve macro-average all datasets and Switch averages FEVER and QAMPARI. This descriptive decoding comparison does not establish decoding-invariant significance; the residual greedy no-op variation of Table S7 bounds nuisance effects but does not enter these comparisons.

| | Change | | Preserve | | Switch | |
|---|---|---|---|---|---|---|
| Method | greedy | $T$=1.00 | greedy | $T$=1.00 | greedy | $T$=1.00 |
| SFT | 44.63 | 42.94 | 60.95 | 58.92 | 17.48 | 15.94 |
| GRPO-Marginal | 51.75 | 49.61 | 57.67 | 56.31 | 20.04 | 17.81 |
| Endpoint-LOO | 52.34 | 50.72 | 60.29 | 58.74 | 20.65 | 18.34 |
| CRPO-Evidence | 56.96 | 55.21 | 65.92 | 64.35 | 29.15 | 26.69 |
| Typed-Pair | 55.81 | 54.19 | 64.62 | 62.47 | 27.94 | 25.21 |
| PTPO-Shared | 56.48 | 54.63 | 65.46 | 63.98 | 29.73 | 27.41 |
| **PTPO** | 57.75 | 56.02 | 66.58 | 65.24 | 33.42 | 30.14 |

Table S25: Generalization checks (five-seed mean ± SD). (a) All RL curricula exclude Forced Switch edges; strict curricula additionally remove natural-redundancy edges together with their dedicated alternative-support nodes; the full-curriculum row is PTPO's main result. Exclusions act at the root level on the RL graph, with Stage-A shared across methods (Table S18). (b) Held-out support topologies; the $\geq 3$ row holds out the per-target predicate, not the edge-level family count of Table S11(c). (c) MuSiQue-to-QAMPARI transfer under the unified static direct-context interface; the in-domain reference row uses the same interface, and target-data exclusion applies only to the transfer row. The retrieval-agent condition is separate (Table S14). (d) Confirmation on a second backbone, not cross-model transfer.

**(a) Curricula without Forced Switch edges (complete results of Figure 6)**

| Training exposure | FEVER Switch | QAMPARI Switch | Macro |
|---|---|---|---|
| SFT (start; shared Stage-A) | 19.69 ±0.39 | 15.27 ±0.49 | 17.48 ±0.38 |
| GRPO-Marginal, strict | 21.44 ±0.42 | 16.85 ±0.55 | 19.14 ±0.41 |
| CRPO-Evidence, strict | 27.81 ±0.51 | 23.15 ±0.68 | 25.48 ±0.55 |
| **PTPO, strict** | 31.44 ±0.48 | 26.55 ±0.62 | 28.99 ±0.51 |
| **PTPO, full curriculum** | 35.94 ±0.50 | 30.91 ±0.82 | 33.42 ±0.38 |

**(b) Held-out support topologies**

| Held-out topology | Edges | SFT Change | PTPO Change | SFT Preserve | PTPO Preserve |
|---|---|---|---|---|---|
| Joint multi-sentence proof | 214 | 40.23 ±0.41 | 53.12 ±0.47 | 55.83 ±0.38 | 60.41 ±0.44 |
| Hop $\geq 4$ | 186 | 31.62 ±0.45 | 44.92 ±0.52 | 48.21 ±0.41 | 55.13 ±0.48 |

| Held-out topology | Edges | SFT Switch | PTPO Switch |
|---|---|---|---|
| $\geq 3$ alternative proofs | 128 | 14.22 ±0.42 | 28.59 ±0.55 |

**(c) Cross-dataset transfer (unified static direct-context interface)**

| Method | Endpoint | Change | Preserve | Switch |
|---|---|---|---|---|
| SFT (QAMPARI in-domain) | 26.42 ±0.31 | 37.81 ±0.38 | 58.13 ±0.33 | 15.27 ±0.49 |
| **PTPO (transfer: MuSiQue → QAMPARI)** | 28.12 ±0.36 | 41.32 ±0.52 | 60.23 ±0.45 | 21.45 ±0.66 |
| **PTPO (QAMPARI in-domain ref.)** | 30.28 ±0.41 | 50.41 ±0.71 | 64.38 ±0.63 | 30.91 ±0.82 |

**(d) Second backbone: Llama-3.1-8B-Instruct (confirmation only)**

| Method | MuSiQue Endpoint | MuSiQue Change | MuSiQue Preserve | FEVER Switch | QAMPARI Switch |
|---|---|---|---|---|---|
| SFT | 37.42 ±0.35 | 46.83 ±0.40 | 60.12 ±0.38 | 21.31 ±0.45 | 17.09 ±0.52 |
| GRPO-Marginal | 43.12 ±0.38 | 54.62 ±0.48 | 57.21 ±0.42 | 24.13 ±0.58 | 19.15 ±0.68 |
| **PTPO** | 43.03 ±0.36 | 59.32 ±0.45 | 65.42 ±0.40 | 37.19 ±0.55 | 32.36 ±0.62 |

Table S26: Mechanism comparison (five-seed mean ± SD). Configuration and outcome rows align. The arrow rows separately change the kernel and credit convention. Typed-Pair with independent sampling also changes sampling and rollout reuse; it is not a pairing-only ablation. Budget controls are in Table S8. This is not a complete factorial design.

**(a) Configuration**

| Configuration | Kernel | Pairing | Credit | Reuse |
|---|---|---|---|---|
| CRPO-Evidence | binary | full | joint-std | shared |
| → typed kernel | typed | full | joint-std | shared |
| → LOO (**PTPO**) | typed | full | LOO | shared |
| PTPO-Shared | typed | shared | LOO | shared |
| Typed-Pair (independent) | typed | independent | LOO | independent |
| Binary + Shared (interaction) | binary | shared | LOO | shared |
| $\lambda = 0$ (Endpoint-LOO) | — | — | LOO | shared |

**(b) Outcomes**

| Configuration | MuSiQue Change | MuSiQue Preserve | FEVER Switch | QAMPARI Switch |
|---|---|---|---|---|
| CRPO-Evidence | 56.42 ±0.54 | 62.81 ±0.44 | 31.63 ±0.66 | 26.67 ±0.91 |
| → typed kernel | 56.83 ±0.51 | 63.12 ±0.42 | 33.13 ±0.62 | 28.24 ±0.87 |
| → LOO (**PTPO**) | 57.24 ±0.39 | 63.52 ±0.41 | 35.94 ±0.50 | 30.91 ±0.82 |
| PTPO-Shared | 56.13 ±0.50 | 62.38 ±0.45 | 32.19 ±0.56 | 27.27 ±1.00 |
| Typed-Pair (independent) | 55.42 ±0.53 | 61.62 ±0.39 | 30.31 ±0.59 | 25.58 ±0.94 |
| Binary + Shared (interaction) | 55.41 ±0.55 | 61.72 ±0.47 | 30.13 ±0.61 | 24.97 ±1.02 |
| $\lambda = 0$ (Endpoint-LOO) | 52.87 ±0.42 | 57.64 ±0.32 | 23.13 ±0.55 | 18.18 ±0.75 |

# D COMPLETE DIAGNOSTIC, MATCHED, AND SWITCH-CONTRAST RESULTS

These tables retain the complete numerical record behind the diagnostic figure and the matched-endpoint comparison (the curriculum results are in Table S25(a)), together with the six Switch contrasts and their $p$-values. Their layout and emphasis follow the main training table.

Table S27: Behavioral diagnosis: endpoint reinforcement raises every semantic endpoint while lowering preservation. Base is a fixed-prompt reference with no training-seed spread. The lower block reports GRPO-Marginal minus SFT; complete absolute scores for both trained methods are in Table 1. Change/Preserve are root/type-balanced F1 and Switch uses a fixed eligible-edge denominator whose macro excludes MuSiQue. Endpoint is task-native. Statistical inference uses root-clustered intervals, not seed SD.

| Method | Endpoint | Change | Preserve | Switch |
|---|---|---|---|---|
| *Base (fixed-prompt reference)* | | | | |
| MuSiQue-Full | 24.63 | 28.41 | 41.62 | — |
| FEVER | 68.31 | 33.24 | 52.68 | 8.44 |
| QAMPARI | 17.93 | 24.12 | 44.31 | 6.06 |
| *Macro* | — | 28.59 | 46.20 | 7.25 |
| *Endpoint reinforcement: GRPO-Marginal − SFT* | | | | |
| MuSiQue-Full | +5.68 | +7.62 | −3.18 | — |
| FEVER | +4.68 | +7.34 | −3.17 | +2.81 |
| QAMPARI | +3.71 | +6.40 | −3.51 | +2.30 |
| *Macro* | — | +7.12 | −3.28 | +2.56 |

Table S28: Full-test results after development-set endpoint matching. Five-seed mean ± sample SD. Checkpoints are chosen independently per seed within predeclared development ranges and frozen before test evaluation (Table S23). Matching constrains aggregate endpoints, not optimization history or instance-level errors. QAMPARI uses static direct context; agent results are separate (Table S14).

| Dataset | Method | Test Endpoint | Change | Preserve | Switch |
|---|---|---|---|---|---|
| MuSiQue-Full | GRPO-Marginal | 40.61 ±0.35 | 52.13 ±0.42 | 55.42 ±0.39 | — |
| | Endpoint-LOO | 40.32 ±0.29 | 52.74 ±0.38 | 57.83 ±0.35 | — |
| | CRPO-Evidence | 40.12 ±0.28 | 56.03 ±0.51 | 62.41 ±0.47 | — |
| | **PTPO** | 40.72 ±0.34 | 56.91 ±0.43 | 63.14 ±0.38 | — |
| FEVER | GRPO-Marginal | 83.52 ±0.27 | 58.53 ±0.37 | 63.02 ±0.38 | 22.31 ±0.61 |
| | CRPO-Evidence | 83.83 ±0.31 | 64.52 ±0.46 | 71.03 ±0.42 | 31.56 ±0.63 |
| | **PTPO** | 83.91 ±0.33 | 65.42 ±0.39 | 71.62 ±0.39 | 33.00 ±0.53 |
| QAMPARI | GRPO-Marginal | 30.02 ±0.47 | 44.03 ±0.61 | 54.41 ±0.63 | 17.33 ±0.87 |
| | CRPO-Evidence | 29.53 ±0.51 | 49.42 ±0.73 | 63.53 ±0.59 | 26.30 ±0.94 |
| | **PTPO** | 30.21 ±0.44 | 50.22 ±0.68 | 64.22 ±0.60 | 30.67 ±0.85 |

Table S29: **Switch contrasts with inferential uncertainty.** PTPO minus each comparison method on macro SwitchAcc (FEVER and QAMPARI), in the row order of Table 1; the strongest prior-method baseline, CRPO-Evidence, is shaded. Intervals are 95% paired root-clustered bootstrap intervals of five-seed means (Table S3(b)) and describe inferential uncertainty, not training-seed spread. All six contrasts belong to the pre-registered Switch family; $p$-values are raw and Holm-corrected (Table S17). The $p$-values are two-sided Wald values derived from the same paired bootstrap standard errors that set the interval widths, not independent resampling evidence (Appendix A.8).

| Comparison | $\Delta$ SwitchAcc | 95% interval | $p$ (raw) | $p$ (Holm) |
|---|---|---|---|---|
| PTPO vs SFT | +15.94 | [12.20, 19.68] | $< 10^{-16}$ | $< 10^{-16}$ |
| PTPO vs GRPO-Marginal | +13.39 | [9.90, 16.88] | $5.60 \times 10^{-14}$ | $2.24 \times 10^{-13}$ |
| PTPO vs CRPO-Evidence | +4.28 | [2.21, 6.35] | $5.12 \times 10^{-5}$ | $1.02 \times 10^{-4}$ |
| PTPO vs Endpoint-LOO | +12.77 | [9.62, 15.92] | $1.78 \times 10^{-15}$ | $8.88 \times 10^{-15}$ |
| PTPO vs Typed-Pair | +5.48 | [3.05, 7.91] | $9.90 \times 10^{-6}$ | $2.97 \times 10^{-5}$ |
| PTPO vs PTPO-Shared | +3.69 | [1.42, 5.96] | $1.43 \times 10^{-3}$ | $1.43 \times 10^{-3}$ |