REWARDING THE GRAPH BEHIND THE CHAIN: TOPOLOGY-AWARE PROCESS SUPERVISION FOR RL AND REASONING DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form reasoning is produced as a linear token sequence, even when a conclusion depends on multiple non-adjacent premises. Outcome-only rewards do not expose this support structure, while scalar process scores do not explicitly identify which earlier premises support a later conclusion. We introduce TopoPRM, a hierarchical process reward model that recovers typed support graphs from ordinary reasoning text. The same dependency representation drives two forms of supervision: asymmetric credit estimation (ACE) assigns structural credit under final-answer correctness constraints, and topology-guided distillation (TGD) directs a fixed teacher to revise the current student's reasoning. Fresh student rollouts close the loop between structural diagnosis and online supervision. The deployed policy generates ordinary text without graph decoding or a verifier. Across four mathematics benchmarks, the complete pipeline improves pass@1 by up to 13.3 percentage points over outcome-only training while reducing mean response length by 3.8%–24.0%. Against the distillation-matched *w/o ACE* control, it gains 2.6 points in average pass@1 and uses 14.8% fewer tokens. These results connect correctness-constrained structural credit with more compact reasoning, using a shared representation for reward and targeted revision. The source code is available at `https://anonymous.4open.science/r/TopoPRM-B47E/`.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) now drives long chain-of-thought (CoT) reasoning in large language models and is a core mechanism for training mathematical and symbolic ability (Wei et al., 2022; Kojima et al., 2022; DeepSeek-AI, 2025; Yeo et al., 2025; Shao et al., 2024; Ke et al., 2025). From sparse final-answer rewards (Cobbe et al., 2021) to process reward models (PRMs) that attach supervision to intermediate positions or prefixes (Lightman et al., 2023; Wang et al., 2024; Uesato et al., 2022), the dominant training interface still treats a trace as a *linear sequence of steps*. However, long reasoning is usually expressed sequentially even when its dependencies are non-local: a later conclusion depends on several non-adjacent premises, an intermediate equation is reused downstream, and subgoals branch before merging into the answer (Besta et al., 2024; Zhong et al., 2026). **Closing the structural mismatch** is the core challenge for moving reasoning supervision beyond linear credit assignment.

Recent reasoning models sharpened the macroscopic structure of inference but stayed inside the CoT-as-result view. Outcome-only rewards scale (DeepSeek-AI, 2025; Yu et al., 2025; Lambert et al., 2024), yet they cannot separate a well-supported derivation from one that reaches the same answer through unsupported intermediate claims, and they remain open to reward hacking (Gao et al., 2023; Skalse et al., 2022). PRMs densify the signal with Monte Carlo rollouts, trained verifiers, or implicit outcome-derived scores (Wang et al., 2024; Luo et al., 2024; Zhang et al., 2025; Khalifa et al., 2025; Zhao et al., 2026; Yuan et al., 2025; She et al., 2025), and can condition on a full prefix. Their scalar outputs leave the support relations between individual claims implicit. We ask how these relations can guide both process credit and teacher revision.

We introduce TopoPRM, a hierarchical process reward model that uses implicit dependencies in ordinary reasoning text to guide post-training. A frozen extractor builds a typed support DAG,

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.

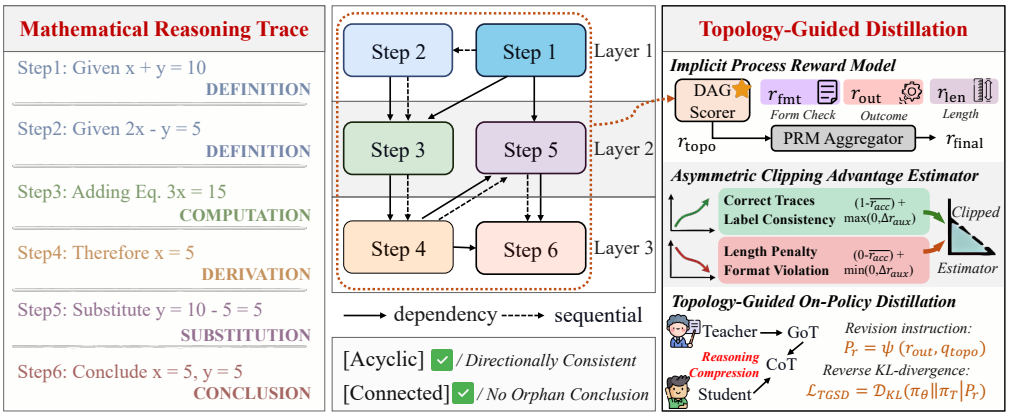

Figure 1: TopoPRM overview. The mathematical trace is mapped to a layered dependency graph. Direction and acyclicity are construction guards, which enter the topology-conditioned multi-source reward. The same graph diagnostics localize dependency-preserving teacher revisions for distillation.

connecting intermediate claims to recoverable premises. TopoPRM combines these dependencies with local continuity in an outcome-anchored reward. Asymmetric credit estimation (ACE) then assigns structural credit within correctness strata, while topology-guided distillation (TGD) uses the same diagnostics to direct teacher revisions of current student rollouts. The shared dependency representation connects reward-based optimization with targeted supervision, while preserving ordinary text generation at inference. Our contributions are threefold:

1. **Implicit dependencies as process supervision.** We propose TopoPRM, which recovers typed support graphs from ordinary reasoning text and combines conclusion support with local continuity in a hierarchical reward, preserving the policy's standard text interface.
2. **Correctness-constrained credit and topology-guided revision.** We propose ACE, which uses within-stratum structural differences to assign asymmetric credit, including when scalar rewards tie. TGD turns localized dependency defects into teacher revision requests on fresh student rollouts, reusing the graph for online supervision.
3. **Accuracy–efficiency and audit evidence.** Across three post-training lineages, Full attains the highest displayed macro-average in every block. Controlled comparisons, training trajectories, edge audits and cross-backbone stress tests characterize when the gains hold and where the surface extractor fails.

2 METHOD

A correct answer leaves open which intermediate claims supported it and which parts of the derivation should be reinforced. TopoPRM uses implicit dependencies in reasoning text to answer both questions: a hierarchical process reward evaluates the trace, ACE assigns structural credit under correctness constraints, and TGD turns localized defects into teacher supervision (Figure 2). Stage I supplies a supervised initialization. Stages II and III refine the student through reinforcement learning and online distillation. Both stages use the same frozen extractor.

2.1 TOPOLOGY-AWARE REASONING REPRESENTATION

A sequential trace can reuse an early equation in several later deductions or combine distant premises in one conclusion. To expose these relations, let $y \sim \pi_\theta(\cdot \mid x)$ for a question and optional context $x = (q, c)$. The frozen extractor segments y into steps $(s_1, \ldots, s_T)$ and returns $\mathcal{E}(y) = (\mathcal{G}_y, \mathcal{L}_y)$: a typed support DAG and a ledger recording the cues that trigger each edge. The DAG aggregates structural quality, while the ledger retains the locations needed for revision.

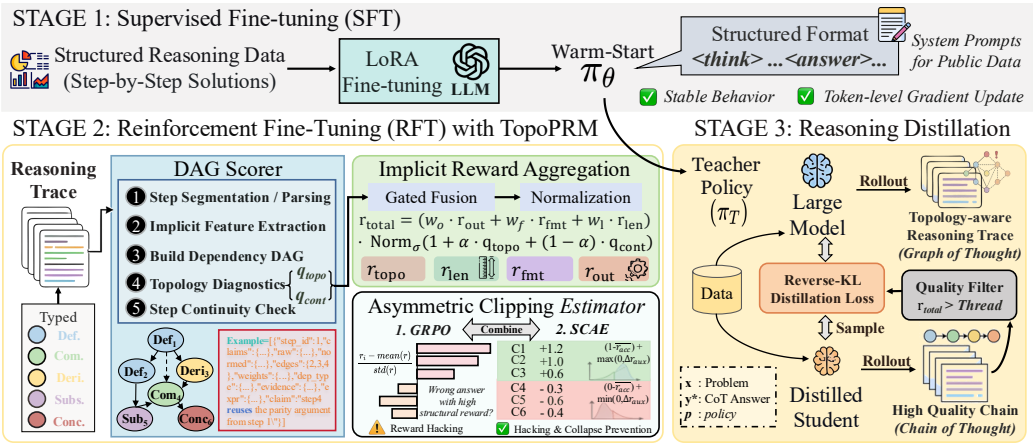

Figure 2: TopoPRM pipeline. A shared dependency representation connects hierarchical process rewards, asymmetric credit estimation (ACE), and topology-guided distillation (TGD), supplying rollout scores and step locations for teacher feedback. Deployment uses the student alone.

Typed dependency construction. Within each subproblem, an edge $i \to j$, with $i < j$, connects steps through expression reuse, claim reference, expression overlap, or guarded variable reuse. A derivation or conclusion lacking stronger support may receive a lower-weight implicit-block edge from s_{j-1}; eligible adjacent steps may also receive a sequential-fallback edge. Each record stores the endpoints, relation type, triggering spans, and weight $\omega_{ij} \in [0, 1]$ (Appendix D). This typed representation distinguishes symbolic reuse from weaker order-based links even when both contribute to the same aggregate score.

The forward restriction enforces direction consistency and acyclicity. The variable training signals are conclusion support and stepwise traceability; direction and cycle checks diagnose graph validity, and reference-edge agreement is evaluated separately in the offline audit. Thus, the graph supplies a stable structural interface without requiring the student to emit graph syntax.

Complementary structural diagnostics. The graph exposes conclusion support and grounding throughout a derivation. For conclusion nodes $\mathcal{C}_y$, let $\sigma_j = \min\{1, \sum_{i:i\to j} \omega_{ij}\}$ and $\rho_{\text{orph}} = |\mathcal{C}_y|^{-1} \sum_{j \in \mathcal{C}_y} (1 - \sigma_j)$, with $\rho_{\text{orph}} = 0$ for an empty conclusion set. An empty graph receives $q_{\text{topo}} = 0$. For a nonempty graph, we separate validity from conclusion support:

$$g_{\text{valid}}(y) = \lambda_b \mathbb{1}[|\mathcal{V}_y| > 0] + \lambda_a \mathbb{1}[\mathcal{G}_y \text{ is acyclic}] + \lambda_d \delta(\mathcal{G}_y),$$
$$q_{\text{topo}}(y) = \frac{g_{\text{valid}}(y) + \lambda_o \mathbb{1}[\rho_{\text{orph}} = 0] + \lambda_k \kappa_{\text{edge}}(\mathcal{G}_y, \mathcal{G}^\star)}{\lambda_b + \lambda_a + \lambda_d + \lambda_o + \lambda_k}. \tag{1}$$

Here δ measures forward-edge consistency and κ_{edge} is edge F1 against an aligned reference. Rollouts use $\lambda_k = 0$. Direction and acyclicity are fixed by forward construction, while graph existence and conclusion support can vary. The node deficits $1 - \sigma_j$ identify locations for TGD, so the aggregate reward and revision request use the same extracted relations.

Continuity evaluates grounding across all parsed steps. Let $h_i = 1$ if s_i explicitly refers to the problem or reuses an extracted expression or claim from an earlier step, and $h_i = 0$ otherwise. A step with neither an extractable expression nor a claim receives the parser fallback $h_i = 1$. For $T > 0$, write $\eta(y) = T^{-1} \sum_i h_i$. We use $q_{\text{cont}} = 1$ when every step is grounded and $q_{\text{cont}} = 0.8\eta$ otherwise. An empty trace receives zero. The fallback and the all-grounded bonus are specified in Appendix D.

Topology measures conclusion support, including distant dependencies, while continuity measures grounding throughout the derivation. A conclusion can reuse an early equation even when an intervening step lacks grounding. The two diagnostics retain separate scores and revision locations for these cases.

2.2 CORRECTNESS-CONSTRAINED STRUCTURAL OPTIMIZATION

Structural quality must influence learning without overriding answer correctness. A bounded scalar reward can also clip away differences among successful traces. Hierarchical aggregation and ACE address these problems at different levels: the former scores rollouts; the latter assigns correctness-constrained structural credit that remains informative after reward saturation.

Outcome-anchored hierarchical reward. The base reward combines binary correctness $r_{\text{out}} \in \{0, 1\}$ with format and length scores $r_{\text{fmt}}, r_{\text{len}} \in [0, 1]$:

$$r_{\text{base}} = w_o r_{\text{out}} + w_f r_{\text{fmt}} + w_l r_{\text{len}}, \qquad w_o + w_f + w_l = 1, \quad w_o, w_f, w_l \geq 0. \tag{2}$$

Within a GRPO group, each structural component is min–max normalized to $[0, 1]$, with a constant component mapped to $1/2$. The structural mixture u modulates the base reward:

$$\begin{aligned} u &= \alpha_{\text{mix}} \widetilde{q}_{\text{topo}} + (1 - \alpha_{\text{mix}}) \widetilde{q}_{\text{cont}}, \\ r_{\text{total}} &= \text{clip}_{[0,1]}(\max(r_{\text{base}}, \epsilon)(1 + u)), \end{aligned} \tag{3}$$

where $\alpha_{\text{mix}} \in [0, 1]$ and $\epsilon > 0$ is the base floor. Multiplication scales the structural increment with the base reward; clipping bounds the resulting signal. With two rollouts per prompt, min–max normalization preserves the ordering of unequal component scores. ACE uses u before the final clipping, retaining structural distinctions that the bounded scalar reward may lose.

Asymmetric credit estimation (ACE). ACE adapts correctness-stratified auxiliary credit (Liu et al., 2026) to the bounded hierarchical reward. It combines a standardized total-reward baseline with structural deviations computed before clipping, then projects the coefficient onto the interval allowed by answer correctness. Let z_i be rollout i's group-standardized total reward, and let $\mathcal{S}_i$ contain rollouts with the same correctness label. Its structural deviation is $\Delta u_i = u_i - |\mathcal{S}_i|^{-1} \sum_{j \in \mathcal{S}_i} u_j$. The scalar baseline uses the entire prompt group. The structural baseline compares traces within their correctness stratum. Constant scalar rewards give $z_i = 0$, and singleton strata give $\Delta u_i = 0$.

With $\mathcal{C} = \{i : r_{\text{out}}^{(i)} = 1\}$, $\mathcal{W} = \{i : r_{\text{out}}^{(i)} = 0\}$, and bounds $\underline{c} \leq 0 \leq \bar{c}$, the rollout coefficient is

$$\widehat{A}_i = \begin{cases} \text{clip}(z_i + \max(0, \Delta u_i), 0, \bar{c}), & i \in \mathcal{C}, \\ \text{clip}(z_i + \min(0, \Delta u_i), \underline{c}, 0), & i \in \mathcal{W}. \end{cases} \tag{4}$$

Stage II substitutes $\widehat{A}_i$ for the rollout advantage in GRPO (Shao et al., 2024). Correctness determines the admissible sign, and relative structural quality determines the one-sided adjustment. The *w/o ACE* control instead uses z_i, retaining the hierarchical reward and subsequent TGD stage.

Proposition 1 (Sign separation and structural credit under ties)**.** *ACE gives $\widehat{A}_i \in [0, \bar{c}]$ for correct rollouts and $\widehat{A}_i \in [\underline{c}, 0]$ for incorrect rollouts. For a prompt group consisting of exactly two correct rollouts ($G = 2$), with equal scalar rewards and $u_1 > u_2$,*

$$\widehat{A}_1 = \min\left(\bar{c}, \frac{u_1 - u_2}{2}\right), \qquad \widehat{A}_2 = 0.$$

The tie case explains why structural credit is retained separately from the bounded reward. Even when successful traces receive identical scalar rewards, ACE can reinforce the trace with stronger extracted support. With one correct and one incorrect rollout, both strata are singletons and the structural deviations vanish. ACE then acts through sign clipping alone. Appendix B.3 derives both cases and the clipping threshold.

2.3 TOPOLOGY-GUIDED ONLINE DISTILLATION

A structural reward says how strongly to reinforce a trace but does not show the student how to repair a missing connection. TGD converts the same diagnostics into targeted supervision. At iteration k, the current student π_{θ_k} samples $y^\star$ for input x. An orphan conclusion or continuity break determines a revision instruction P_r containing the defect type and localized step. A fixed teacher π_T revises the trace conditioned on $(x, y^\star, P_r)$. After updating the student, fresh rollouts produce the next revision requests, keeping supervision tied to its current reasoning behavior.

The student starts from Stage II; the fixed teacher has the same architecture and parameter count. When no structural defect is localized, the request targets answer correction or concise restatement. The gate a_r accepts $\widetilde{y}$ when answer and format checks pass, neither structural score decreases relative to $y^\star$, and the response meets the prescribed budget (Appendix B.4).

Gated distillation objective. At a revised prefix $\widetilde{y}_{<t}$, let $p_t = \pi_\theta(\cdot \mid x, \widetilde{y}_{<t})$ and $q_t = \pi_T(\cdot \mid x, y^\star, P_r, \widetilde{y}_{<t})$. The student minimizes

$$\mathcal{L}_{\mathrm{TGD}}^{(k)}(\theta) = \mathbb{E}_{\substack{x,\, y^\star \sim \pi_{\theta_k}(\cdot|x) \\ \widetilde{y} \sim \pi_T(\cdot|x, y^\star, P_r)}} \left[a_r(x, y^\star, \widetilde{y}) \sum_{t=1}^{|\widetilde{y}|} \mathcal{D}_{\mathrm{RKL}}(p_t \,\|\, q_t) \right], \tag{5}$$

where $\mathcal{D}_{\mathrm{RKL}}(p_t \,\|\, q_t) = \mathrm{KL}(p_t \,\|\, q_t)$. Sampling, acceptance decisions, and teacher distributions are fixed during each gradient update. Initial traces follow the current student, while token supervision uses teacher-revised prefixes. The teacher receives the original trace and revision instruction, while the student learns the revised reasoning from the task context alone.

The graph therefore links two forms of supervision: rollout-level structural credit and step-localized teacher feedback. Extraction considers at most $O(T^2)$ candidate step pairs after segmentation. Graph extraction and teacher calls occur during training. The deployed student produces ordinary text.

3 Experiments

The evaluation addresses four questions. RQ1 asks whether gains recur across model lineages and where Full lies on the accuracy–length frontier. RQ2 examines sensitivity to credit constraints, topology, and continuity. RQ3 tests extracted support against separately produced references, and RQ4 asks whether topology adds ranking information beyond a semantic PRM. Appendix E provides repeated-decoding, structure–correctness, cross-backbone, and shared-pool tests.

Table 1: Pass@5 accuracy (%) on nine public benchmarks. The upper block provides complete open-checkpoint references; the lower blocks compare post-training methods within the same model family. **Bold** and underlined values denote the best and second-best result within each post-training block, respectively; tied best results are bold. Avg. is the reported nine-task macro-average. All values are preserved from the archived five-sample result ledger; Appendix C separates this protocol from the controlled pass@1 study and records the provenance scope.

Model	Params	Elementary		Olympiad		Competition			General		Avg.
		GSM8K	MATH	Olympiad	Omni	AIME'24	AIME'25	CNMO'24	MMLU	GPQA-D	
Complete open checkpoints											
Qwen3.5-2B	2B	77.2	60.8	48.5	13.3	6.7	13.3	36.1	32.8	37.9	36.3
Qwen3.5-4B	4B	85.1	55.8	32.8	43.4	0.0	13.3	10.0	67.4	29.8	37.5
Qwen2.5-Math-7B-Instruct	7B	95.2	73.8	38.0	22.0	13.0	10.0	13.3	54.8	28.3	38.7
Llama-3.1-8B-Instruct	8B	84.5	52.2	15.0	10.0	3.0	2.0	3.0	68.0	30.0	29.7
DeepSeek-R1-Distill-Llama-8B	8B	83.0	89.1	48.0	25.0	50.4	35.0	25.0	52.0	49.0	50.7
Qwen2.5-7B post-training											
Base	7B	85.4	49.8	25.0	18.0	7.0	5.0	8.0	74.2	36.4	34.3
+ SFT	7B	91.3	65.4	42.5	54.2	16.7	12.2	16.7	70.1	28.3	44.2
+ SimpleRL-Zoo	7B	91.7	78.2	38.6	40.4	**20.0**	15.6	15.0	62.5	**55.2**	46.4
+ Open-Reasoner-Zero	7B	89.0	**81.4**	**47.9**	38.4	17.9	15.6	13.3	62.5	36.6	44.7
+ **TopoPRM (Ours)**	7B	**95.8**	72.6	47.8	**57.2**	16.7	**26.7**	**47.0**	**82.7**	51.0	**55.3**
DeepSeek-R1-Distill-Qwen-7B post-training											
Base	7B	73.8	36.8	25.8	28.0	46.7	30.0	22.9	23.7	12.6	33.4
+ SFT	7B	**86.0**	**92.8**	52.0	48.0	**55.5**	**40.0**	30.0	50.0	**49.1**	55.9
+ GRPO	7B	85.1	67.4	56.2	72.0	50.0	**40.0**	**57.8**	50.1	17.7	55.1
+ **TopoPRM (Ours)**	7B	84.3	66.6	**56.4**	**72.2**	**55.5**	**40.0**	55.4	**60.3**	34.3	**58.3**
Qwen3.5-9B post-training											
Base	9B	69.5	66.6	55.2	61.5	16.7	21.1	16.7	28.6	37.4	41.5
+ SFT	9B	94.1	68.2	44.0	53.8	30.0	15.6	30.0	78.4	**47.0**	51.2
+ GRPO	9B	93.3	63.0	**56.2**	72.0	**46.7**	**30.0**	**57.8**	63.0	17.7	55.5
+ **TopoPRM (Ours)**	9B	**97.9**	**82.0**	**56.2**	**76.2**	43.3	18.9	50.0	**86.4**	40.4	**61.3**

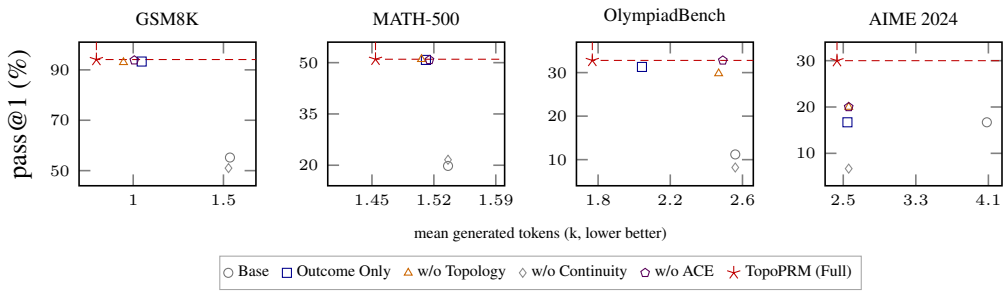

Figure 3: Accuracy–length trade-off among the Qwen3.5-9B checkpoints in Table 2. Full ($\star$) is the only checkpoint on the observed Pareto frontier of all four benchmarks. Axes are cropped per panel; length is in thousands of visible tokens.

Table 2: Qwen3.5-9B ablation from a shared Stage-I initialization. Cells report pass@1 (%) / mean visible tokens, with task averages at right. ✓ indicates TGD. Gold/blue mark w/o ACE/Full. Bold marks best accuracy or shortest responses among post-training variants, including ties.

Configuration	**TGD**	**pass@1↑ / tokens↓**				**Avg.↑**	**Mean↓**
Benchmark		GSM8K	MATH-500	Olympiad-Bench	AIME 2024	pass@1	tokens
Stage-I reference	–	55.2 / 1,536	19.8 / 1,536	11.2 / 2,560	16.7 / 4,083	25.7	2,429
Post-training variants							
Outcome Only	–	93.3 / 1,047	50.8 / 1,511	31.3 / 2,043	16.7 / 2,546	48.0	1,787
w/o Topology	–	93.0 / 946	**51.0** / 1,506	29.8 / 2,468	20.0 / 2,560	48.5	1,870
w/o Continuity	–	51.0 / 1,528	21.6 / 1,536	8.2 / 2,560	6.7 / 2,560	21.9	2,046
w/o ACE	✓	93.8 / 1,006	50.8 / 1,515	**32.8** / 2,491	20.0 / 2,560	49.4	1,893
TopoPRM (Full)	✓	**94.1 / 796**	**51.0 / 1,454**	**32.8 / 1,768**	**30.0 / 2,431**	**52.0**	**1,612**

3.1 Evaluation Setup

Training, baselines, and controls. All controlled variants share a Qwen3.5-9B Stage-I checkpoint (Qwen Team, 2026a) and frozen graph extractor. Stage II applies rank-64 LoRA (Hu et al., 2022) with GRPO (Shao et al., 2024), two completions per prompt, and learning rate 5×10^{-6}. Full and *w/o ACE* proceed to Stage III with a same-architecture, equal-size teacher; *Base* is Stage I, and other controls stop after Stage II.

Table 1 groups complete Qwen3.5, Qwen2.5-Math, Llama 3.1, and DeepSeek-R1-Distill checkpoints (Qwen Team, 2026b;c; Yang et al., 2024; Grattafiori et al., 2024; DeepSeek-AI, 2025) with within-lineage SFT, GRPO, SimpleRL-Zoo (Zeng et al., 2025), and Open-Reasoner-Zero (Hu et al., 2025) baselines. Because external data and compute differ, Table 2 supplies the shared-initialization analysis. *Outcome Only* uses final-answer correctness, and *w/o ACE* retains the structural reward but uses standard group-relative advantages. The topology and continuity controls remove the named source and redistribute its reward mass. The shorter Stage-II training cap for *w/o Continuity* makes the source-removal rows tests of pipeline sensitivity rather than factorial ablations (Appendix B).

Benchmarks and metrics. Table 1 reports five-sample pass@5 across its nine benchmark columns (Cobbe et al., 2021; Hendrycks et al., 2021; He et al., 2024; Gao et al., 2024a; Hendrycks et al., 2020; Rein et al., 2023). The controlled study reports pass@1 and mean visible tokens on GSM8K, MATH-500 (Lightman et al., 2023), a 134-item English text-only OlympiadBench subset, and AIME 2024. Acc/kTok is $\mathrm{pass@1}/(\overline{\mathrm{tokens}}/1000)$ and is not optimized. Figure 4 uses a separate suite; Appendix B records all three protocols.

Each result source is interpreted only within its own protocol: Table 1 measures breadth, Table 2 controls initialization and response cost, and Figure 4 summarizes training behavior.

3.2 Main Results: Accuracy and Efficiency (RQ1)

Observation 1: The complete configuration leads every evaluated post-training lineage in aggregate. TopoPRM reaches macro-averages of 55.3, 58.3, and 61.3 on Qwen2.5-7B, DeepSeek-R1-Distill-Qwen-7B, and Qwen3.5-9B, exceeding the strongest alternative in each block by 8.9, 2.4, and 5.8 points (Table 1). It is best or tied-best on five of nine tasks in every lineage, although the winning tasks vary by backbone. Each TopoPRM row records a best or tied-best result in at least three of the four task categories; the Qwen2.5-7B variant spans all four. The Qwen3.5-9B average of 61.3 is also the table maximum, 10.6 points above the strongest complete-checkpoint reference. This profile is not uniformly dominant: Qwen3.5 GRPO remains stronger on several competition sets, whereas TopoPRM leads that lineage on MATH, Omni-MATH, and MMLU. The evidence therefore supports broad aggregate improvement under the reported protocol, not universal task-wise dominance. For scale context, Qwen2.5-7B TopoPRM (55.3) exceeds the larger DeepSeek-R1-Distill-Llama-8B reference (50.7). Because their data and compute differ, this is contextual rather than a controlled scaling effect; component attribution relies on shared initialization.

Observation 2: Full is the only checkpoint on each observed accuracy–length frontier. Against Outcome Only, Full improves pass@1 by 0.8, 0.2, 1.5, and 13.3 points on GSM8K, MATH-500, OlympiadBench, and AIME 2024 while saving 251, 57, 275, and 115 tokens (Figure 3; Table 2). The easier tasks primarily separate response cost, whereas AIME supplies the largest accuracy gap. Acc/kTok rises from 89.1 to 118.2 on GSM8K and from 6.6 to 12.3 on AIME, with the same direction elsewhere. Full is therefore the only checkpoint on all four observed frontiers.

The two evaluations provide complementary evidence. Table 1 shows that the aggregate gain recurs across the evaluated post-training lineages, while Table 2 shows that Full matches or improves accuracy with shorter responses on all four benchmarks. The observed gains therefore do not require greater verbosity. Relative to the TGD-matched control, Full largely preserves accuracy on the three larger sets while shortening responses, and achieves its largest accuracy gain on AIME 2024 with fewer tokens. These results support a favorable accuracy–length trade-off across the evaluated tasks.

3.3 Ablation Study (RQ2)

Observation 3: ACE improves the accuracy–length trade-off within the complete pipeline; source removals reveal its structural dependence. Full and *w/o ACE* share the hierarchical reward and TGD stage. Full raises average pass@1 from 49.4 to 52.0 while reducing mean response length from 1,893 to 1,612 tokens (Table 2). On GSM8K, MATH-500, and OlympiadBench, it improves or preserves accuracy while removing 210, 61, and 723 tokens. On AIME 2024, it solves three additional problems with 129 fewer tokens. Thus, the observed benefit combines shorter responses at similar accuracy on the larger sets with improved accuracy on AIME. ACE provides a mechanism for this pattern: it can retain within-correctness structural preferences after scalar reward clipping removes them (Proposition 1).

Relative to Full, removing topology lowers OlympiadBench and AIME 2024 accuracy by 3.0 and 10.0 points and adds 700 and 129 tokens. Removing continuity lowers accuracy by 23.3–43.1 points and lengthens responses on every task. Both changes disrupt the intended combination of non-local support and local traceability. These controls also omit TGD and redistribute reward mass; *w/o Continuity* further uses a shorter training cap. Their differences characterize the complete training pipeline, while the TGD-matched comparison above provides the direct ACE contrast.

Training dynamics. In the complementary optimization suite, Full maintains shorter online responses across the plotted updates (Figure 4). The endpoint evaluation shows the same cost ordering: 1,568 versus 1,737 tokens for Outcome Only, with Acc/kTok of 48.1 versus 37.9.

3.4 Auditing Recoverable Support (RQ3)

Audit protocol. Traces are stratified by task and response length. Both model annotators receive identical parsed pairs but neither extractor decisions nor edge types. After independent labeling, the complete set undergoes manual error checking. We report agreement and extractor scores against each model separately, preserving the distinction between annotation stability and extractor–reference agreement without treating the checks as a human gold standard.

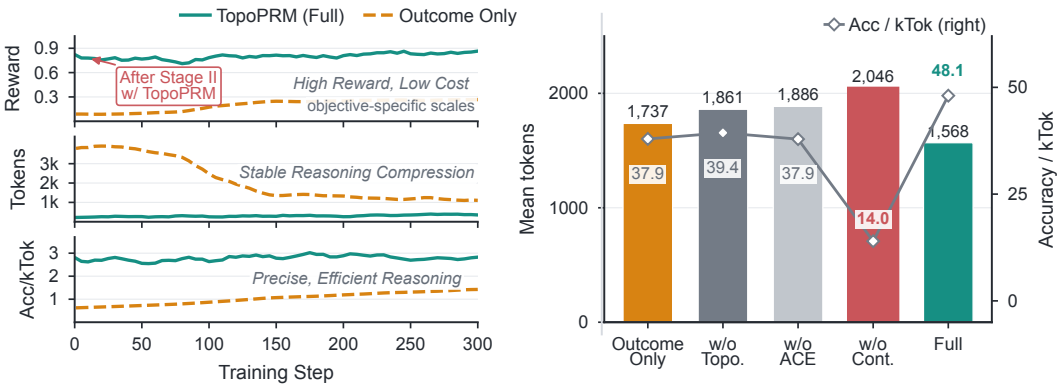

Figure 4: Training and efficiency diagnostics from a complementary run suite. Trajectories summarize reward, response length, and Acc/kTok; bars and diamonds show endpoint length and efficiency. Objective scales differ, and these panels are not pooled with the controlled comparison in Table 2.

Table 3: Edge audit on 1,770 candidate pairs from 119 traces. Qwen3-32B and Qwen2.5-32B label pairs independently. Manual verification covers the complete set. Reported extractor metrics use the model annotations as references. Consensus/union are their positive-label intersection/union.

Annotator contingency
Qwen3-32B vs. Qwen2.5-32B

Qwen3-32B	Support	No support	Total
Support	368	68	436
No support	113	1,221	1,334
Total	481	1,289	1,770

Agreement: raw $= 0.898$, $\kappa = 0.734$

Frozen extractor
against alternative references

Support reference	Prec.	Recall	F1
Qwen3-32B	0.478	0.589	0.528
Qwen2.5-32B	0.483	0.541	0.510
Annotator consensus	0.428	0.625	0.508
Annotator union	0.533	0.523	0.528

Observation 4: Separately produced references are consistent, while extractor–reference agreement remains moderate. Qwen3-32B and Qwen2.5-32B independently labeled 1,770 candidate pairs from 119 traces. They agreed on 89.8%, with Cohen's $\kappa_{\text{ann}} = 0.734$ (Table 3); their positive-label rates are 24.6% and 27.2%, with 368 jointly positive and 181 disputed pairs. Every pair was also checked for evident annotation or parsing errors. Extractor F1 is 0.528 against Qwen3-32B and 0.510 against Qwen2.5-32B; consensus and union references yield 0.508 and 0.528. This narrow range indicates limited sensitivity to the chosen Qwen-family reference, while the moderate absolute F1 precludes a proof-level interpretation. The 120-trace typed ledger attributes 145 false positives to variable-reference edges, compared with 34 expression-reference, 11 expression-overlap, 9 implicit-block, and 82 order-fallback edges. A guarded rule revision raises precision from 0.480 to 0.580 on that ledger. Because the same labels informed its design and evaluation, the result is post-hoc rather than a held-out improvement (Appendix D.4).

3.5 Boundary Analysis with a Semantic PRM (RQ4)

Table 4: Diagnostic PRM–topology fusion on shared Best-of-8 pools ($n = 80$ per benchmark). The hybrid multiplies normalized Qwen2.5-Math-PRM-7B scores by $1 + q_{\text{topo}}$; Δ is the change from PRM selection.

Benchmark	PRM@8	PRM×Topo@8	Δ	Oracle@8
GSM8K	95.00	95.00	0.00	96.25
MATH-500	83.75	83.75	0.00	92.50

Observation 5: Topology does not replace semantic scoring. On shared Best-of-8 pools, multiplying Qwen2.5-Math-PRM-7B scores (Zhang et al., 2025) by $1 + q_{\text{topo}}$ ties the PRM on

GSM8K (95.00%) and MATH-500 (83.75%; Table 4). Topology-only selection trails the PRM on both tasks, while the pool oracle remains 1.25 and 8.75 points higher. The tie bounds the claim: this fusion neither improves aggregate selection nor replaces semantic verification. Establishing complementarity requires larger prespecified pools and item-level disagreement analysis.

3.6 Robustness and Failure Modes

Observation 6: Cross-backbone accuracy is mixed, and stochastic decoding does not preserve the greedy ordering. In the matched DeepSeek-R1-Distill-Qwen-7B study, greedy pass@1 favors Full, but five-draw means span only 67.8–68.0 with standard deviations of 1.8–2.7, so the ordering is not stable under decoding variation. On Llama-3.1-8B-Instruct, Full saves 389 tokens on GSM8K with unchanged pass@1 and 396 tokens on MATH-500 while pass@1 falls from 46.5 to 42.5. Response shortening transfers to these subsets; accuracy ordering and preservation do not (Appendix E).

Observation 7: Topology score is associated with correctness but is not uniquely discriminative. Full has the highest mean topology score (0.939), and its high-score stratum exceeds its low-score stratum by 17.15 accuracy points. Outcome-only training exhibits a comparable 15.04-point gap. The PRM tie and edge-audit F1 likewise support interpreting graph diagnostics as an inspectable structural preference under outcome verification, not as a semantic discriminator.

3.7 Discussion

Table 1 reports aggregate improvements across three post-training lineages. In the shared-initialization suite, Full improves accuracy over Outcome Only while shortening responses on all four benchmarks. The TGD-matched comparison also associates ACE with higher average accuracy and shorter responses (Table 2).

Training analyses show how the structural signals enter learning. ACE retains nonzero credit in 437 of 1,071 groups with tied scalar rewards. At update 177, two correct responses to an orange-bag arithmetic question both receive scalar reward one. Their topology scores are 1 and $10/13$, while both continuity scores are zero. After group normalization, their structural mixtures are 0.8 and 0.2, and ACE assigns coefficients 0.3 and 0. The tied outcome therefore retains a structural learning preference: the better-supported correct response receives positive credit, while the other is not penalized. This recorded pair illustrates Proposition 1 within a training update (Appendix C.3).

In a separate 675-item OlympiadBench comparison, TGD improves accuracy over generic revision by 3.85 percentage points with the same Stage-II initialization, teacher, structural acceptance gate, and revision-attempt budget. TGD solves 169 problems versus 143 for generic revision, with a paired question-bootstrap 95% interval of $[1.19, 6.67]$ percentage points for the accuracy difference (Appendix C.3). The typed-edge audit connects the structural diagnostics to their underlying text relations. The shared dependency representation thus links hierarchical process rewards, correctness-constrained credit, and targeted online teacher revision within TopoPRM. ACE changes how a generated trace is reinforced when scalar rewards tie. TGD changes which dependency defects the teacher is asked to repair before distillation. The graph thus supplies both a preference over student traces and a location for teacher feedback, without requiring graph-formatted policy outputs.

4 Conclusion

TopoPRM makes implicit dependencies in free-form reasoning actionable during post-training. Its hierarchical process reward scores local continuity and non-local support, ACE assigns asymmetric structural credit under answer-correctness constraints, and TGD turns dependency defects in current student rollouts into targeted teacher revisions. Together, these components connect what a reasoning trace should preserve with how the student is reinforced and supervised, without adding graph generation or teacher calls at deployment. Moving beyond surface cues toward semantically grounded dependencies is a promising direction for extending this approach to tool-assisted and multimodal reasoning (Appendix G).

AI USE STATEMENT

Generative AI tools assisted with literature discovery, language editing, LaTeX formatting, methodological and experimental review, mathematical exposition, and interpretation of results. AI-assisted code revision and experiment scripting supported the research workflow. The edge audit uses the model annotators identified in the paper. The authors are responsible for verifying all AI-assisted text, code, analyses, citations, and reported results.

ETHICS STATEMENT

The reported study uses public mathematics benchmarks and model-generated reasoning traces. It collects no personal or demographic data, includes no human-subject study, and uses no private benchmark to support a claim. The edge audit uses model annotators and is identified as such. The principal risk is methodological: a trace can exhibit coherent surface dependencies while containing an incorrect mathematical deduction. We therefore treat topology as a training-time process signal rather than a semantic verifier, report this failure mode directly, and retain final-answer verification as a separate reward source. The study is limited to mathematical reasoning; broader deployment claims would require domain-specific safety and validity evaluation.

REPRODUCIBILITY STATEMENT

Appendix B records the training, reward, and decoding configuration, and Appendix D specifies graph construction and the audit protocol. The anonymized supplementary artifact contains source and configuration snapshots plus aggregate JSON files for the edge audit, controlled rerun, cross-backbone, structural, and shared-pool diagnostics. Table 1 was transcribed from archived aggregate evaluation tables; its per-item outputs are not included, which we disclose as a reproducibility limitation. We distinguish single training runs, repeated decoding draws, and paired annotation statistics throughout, and we do not use incomplete experiments as evidence.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025.

Daman Arora and Andrea Zanette. Training language models to reason efficiently. *Advances in Neural Information Processing Systems*, 38:60770–60808, 2025.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Nyczyk, Marcin Copik, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=lNGCIHWARF.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. Omni-MATH: A universal olympiad-level mathematics benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024a. URL https://arxiv.org/abs/2410.07985.

Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. On designing effective RL reward at training time for LLM reasoning. *arXiv preprint arXiv:2410.15115*, 2024b. URL https://arxiv.org/abs/2410.15115.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Han Xu, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. URL https://arxiv.org/abs/2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 34, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *Advances in Neural Information Processing Systems*, 38:162239–162262, 2025.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, Caiming Xiong, and Shafiq Joty. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems, 2025. URL https://arxiv.org/abs/2504.09037.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think. *arXiv preprint arXiv:2504.16828*, 2025.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Runxuan Liu, Xianhao Ou, Xinyan Ma, Jiyuan Wang, Jiafeng Liang, Jiaqi Li, Tao He, Zheng Chu, Rongchuan Mu, Zekun Wang, et al. Graph reasoning paradigm: Structured and symbolic reasoning with topology-aware reinforcement learning for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026. URL https://aclanthology.org/2026.acl-long.1660/.

Jianqiao Lu, Zhiyang Dou, Hongru Wang, Zeyu Cao, Jianbo Dai, Yingjia Wan, Yunlong Feng, and Zhijiang Guo. Autopsv: Automated process-supervised verifier. *Advances in Neural Information Processing Systems*, 37:79935–79962, 2024.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*, 2025.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024. URL https://arxiv.org/abs/2406.06592.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, 2025. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Sagnik Mukherjee, Abhinav Chinta, Takyoung Kim, Tarun Anoop Sharma, and Dilek Hakkani-Tür. Premise-augmented reasoning chains improve error identification in math reasoning with LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45109–45128, 2025. URL https://proceedings.mlr.press/v267/mukherjee25a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Miao Peng, Nuo Chen, Zongrui Suo, and Jia Li. Rewarding graph reasoning process makes LLMs more generalized reasoners. *arXiv preprint arXiv:2503.00845*, 2025. URL https://arxiv.org/abs/2503.00845.

Qwen Team. Qwen3.5-9B model card. Hugging Face model release, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Qwen Team. Qwen3.5-2B model card. Hugging Face model release, 2026b. URL https://huggingface.co/Qwen/Qwen3.5-2B.

Qwen Team. Qwen3.5-4B model card. Hugging Face model release, 2026c. URL https://huggingface.co/Qwen/Qwen3.5-4B.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. URL https://arxiv.org/abs/2311.12022.

Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shahriari, Sarah Shariari, Sertan Perrin, Abe Friesen, Geoffrey Cideron, Sertan Girgin, Piotr Stanczyk, Andrea Michi, Danila Sinopalnikov, Sabela Ramos, Morgane Riviere, Joaquin Lin, Dylan Hadfield-Menell, Rémi Munos, Olivier Pietquin, and Olivier Bachem. BOND: Aligning LLMs with best-of-N distillation. *arXiv preprint arXiv:2407.14622*, 2024.

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. *arXiv preprint arXiv:2402.03300*, 2024.

Shuaijie She, Junxiao Liu, Yifeng Liu, Jiajun Chen, Xin Huang, and Shujian Huang. R-PRM: Reasoning-driven process reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13438–13451, 2025. URL https://aclanthology.org/2025.emnlp-main.679/.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. PRMBench: A fine-grained and challenging benchmark for process-level reward models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Xiangyu Wen, Junhua Huang, Zeju Li, Min Li, Jianyuan Zhong, Zhijian Xu, Mingxuan Yuan, Yongxiang Huang, and Qiang Xu. Reasoning scaffolding: Distilling the flow of thought from LLMs. In *International Conference on Learning Representations*, pp. 80433–80459, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8248b1ded388fcdbbd121bcdfea3068c-Paper-Conference.pdf.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3351–3363, 2025. doi: 10.18653/v1/2025.emnlp-main.165. URL https://aclanthology.org/2025.emnlp-main.165/.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. KDRL: Post-training reasoning LLMs via unified knowledge distillation and reinforcement learning. *arXiv preprint arXiv:2506.02208*, 2025. URL https://arxiv.org/abs/2506.02208.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

Edward Yeo, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in LLMs. *arXiv preprint arXiv:2502.03373*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Hongyuan Yuan, Xinran He, Run Shao, Bolei He, Xianwei Xue, Mengke Chen, Qiutong Pan, Haiwei Wang, and Haifeng Li. Graph-based chain-of-thought pruning for reducing redundant reflections in reasoning LLMs. *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. URL https://arxiv.org/abs/2604.05643.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73511–73525, 2025. URL https://proceedings.mlr.press/v267/yuan25c.html.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. SimpleRL-Zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025. URL https://arxiv.org/abs/2503.18892.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772, 2024.

Yuanhe Zhang, Ilja Kuzborskij, Jason Lee, Chenlei Leng, and Fanghui Liu. DAG-Math: Graph-of-thought guided mathematical reasoning in LLMs. In *International Conference on Learning Representations*, pp. 74347–74389, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/79179a9c80561608fa85f925cc88aaa9-Paper-Conference.pdf.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, and Bowen Zhou. GenPRM: Scaling test-time compute of process reward models via generative reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34932–34940, 2026. doi: 10.1609/aaai.v40i41.40797. URL https://ojs.aaai.org/index.php/AAAI/article/view/40797.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning, 2024. URL https://arxiv.org/abs/2412.06559.

Tianjun Zhong, Linyang He, Ziyang Li, and Nima Mesgarani. From chains to dags: Probing the graph structure of reasoning in llms. *arXiv preprint arXiv:2601.17593*, 2026. URL https://arxiv.org/abs/2601.17593.

A Related Work

Reward supervision for reasoning. RLHF learns preference rewards (Ouyang et al., 2022), whereas RLVR uses programmatic answer checks (Shao et al., 2024; DeepSeek-AI, 2025; Lambert et al., 2024; Zeng et al., 2025); subsequent objectives improve optimization and credit propagation (Yu et al., 2025; Kazemnejad et al., 2025). Terminal rewards, however, cannot distinguish differently supported traces with the same answer and remain vulnerable to gaming (Gao et al., 2023; Skalse et al., 2022; Gao et al., 2024b). PRMs provide denser feedback through human labels (Uesato et al., 2022; Lightman et al., 2023), rollout or search (Wang et al., 2024; Luo et al., 2024; Zhang et al., 2024), induced annotations (Lu et al., 2024), outcome-derived values (Yuan et al., 2025; Cui et al., 2026), or generative verification (Khalifa et al., 2025; Zhao et al., 2026); PRMBench and ProcessBench nevertheless reveal persistent error-detection and generalization failures (Song et al., 2025; Zheng et al., 2024). Scalar completion, step, or prefix scores summarize reasoning quality without explicitly identifying the non-adjacent premises supporting a conclusion. *TopoPRM complements answer verification with auditable local-continuity and non-local-support signals extracted from free text.*

Structured reasoning representations. ToTs and GoTs expose branching search and dependencies (Yao et al., 2023; Besta et al., 2024); PARC represents explicit premise links (Mukherjee et al., 2025), while DAG-Math uses graph-structured mathematical reasoning (Zhang et al., 2026). GraphPRM learns process rewards from graph-algorithm trajectories (Peng et al., 2025). Correctness-stratified clipping has also been studied for graph-formatted reasoning (Liu et al., 2026). *TopoPRM instead extracts evidence-bearing support from ordinary reasoning text and uses it for training-time credit assignment and targeted revision, without changing the policy output or inference interface.*

Reasoning distillation and compression. Knowledge distillation transfers teacher behavior (Hinton et al., 2015); language-model variants address exposure bias through reverse-KL, on-policy, adaptive off-policy, Best-of-N, black-box, or reward-coupled objectives (Gu et al., 2024; Agarwal et al., 2024; Ko et al., 2024; Sessa et al., 2024; Xu et al., 2025; Ye et al., 2025; Yang et al., 2026). Reasoning compression controls budgets, prunes tokens, tunes for length, or shapes efficiency rewards (Snell et al., 2024; Muennighoff et al., 2025; Xia et al., 2025; Luo et al., 2025; Arora & Zanette, 2025; Aggarwal & Welleck, 2025); structured variants prune reasoning DAGs or distill semantic operations (Yuan et al., 2026; Wen et al., 2026). *TopoPRM uses defects detected in the student's own trace to target teacher revision, while answer, structure, and response budget jointly determine which revisions become training targets.*

B Implementation and Evaluation Details

B.1 Training Configuration

Table 5 documents the shared SFT and GRPO configuration for the controlled Qwen3.5-9B suite. All variants use the same Stage-I initialization and frozen DAG Scorer. Full and *w/o ACE* additionally use the same TGD procedure and fixed same-parameter teacher checkpoint. Source-removal controls redistribute reward mass, and *w/o Continuity* uses a shorter completion cap; these rows are matched at initialization but are not factorial.

Table 5: Training configuration for the controlled Qwen3.5-9B runs.

Parameter	**Stage I: SFT**	**Stage II: GRPO**
Initialization	Qwen3.5-9B	Stage-I checkpoint
Adapter	LoRA $r = 64$, $\alpha = 128$	LoRA $r = 64$, $\alpha = 128$
Learning rate	5×10^{-5}	5×10^{-6}
Schedule / warmup	cosine / 3%	cosine / 5%
Training length	2 epochs	1 epoch
Microbatch / accumulation	2 / 8	1 / 16
Completions / prompt	–	2
Sampling / nucleus	–	0.8 / 0.95
Completion cap	–	4,096; 1,024 for w/o continuity
Maximum sequence	4,096	8,192; 4,096 for w/o continuity
Precision	bf16	bf16

B.2 Reward Configuration

Equation 1 uses $(\lambda_b, \lambda_a, \lambda_d, \lambda_o, \lambda_k) = (0.20, 0.15, 0.15, 0.15, 0.25)$. For online rollouts, $\lambda_k = 0$ and the remaining weights sum to 0.65. Under forward-only construction, a nonempty graph receives $q_{\text{topo}} = 1$ when all conclusions have saturated support and $10/13$ otherwise. Empty graphs receive zero. The graded node deficits $1 - \sigma_j$ retain the locations used for teacher revision.

The hierarchical base uses $(w_o, w_f, w_l) = (0.70, 0.15, 0.15)$, $\alpha_{\text{mix}} = 0.60$, and $\epsilon = 0.05$. Outcome-only uses r_{out}. The no-topology and no-continuity controls assign $(0.55, 0.15, 0.20, 0.10)$ to outcome, format, the retained structural term, and length. Thus, a removed source is replaced rather than simply set to zero in Eq. 3. The length score is one through 2,000 decoded characters, decreases linearly, and reaches zero at 4,000; all benchmarks use the same constants.

A structural component that is constant within a rollout group is mapped to $1/2$ before Eq. 3. Gaussian variance injection and dynamic reweighting are disabled. The reward and extractor stay fixed within each training run.

B.3 ACE: Derivation and Boundary Cases

For a prompt group of $G > 1$ rollouts, let $r_i = r^{(i)}_{\text{total}}$. The standardized scalar return and the correctness-centered structural deviation are

$$\begin{aligned} &\bar{r} = G^{-1} \sum_j r_j, && s_r^2 = (G-1)^{-1} \sum_j (r_j - \bar{r})^2, && z_i = \frac{r_i - \bar{r}}{s_r + 10^{-4}}, \\ &u_i = \alpha_{\text{mix}} \widetilde{q}^{(i)}_{\text{topo}} + (1 - \alpha_{\text{mix}}) \widetilde{q}^{(i)}_{\text{cont}}, && \bar{u}_{\mathcal{S}_i} = |\mathcal{S}_i|^{-1} \sum_{j \in \mathcal{S}_i} u_j, && \Delta u_i = u_i - \bar{u}_{\mathcal{S}_i}. \end{aligned} \tag{6}$$

The two baselines serve different purposes. Standardization compares overall returns across the prompt group, while $\bar{u}_{\mathcal{S}_i}$ compares structure among traces sharing an answer-correctness label. Constant rewards give $z_i = 0$; singleton strata give $\Delta u_i = 0$. An empty stratum contains no rollout requiring an update. The *w/o ACE* control substitutes z_i for $\widehat{A}_i$ in the same training pipeline.

Proof of Proposition 1. For a correct rollout, clipping projects $z_i + \max(0, \Delta u_i)$ onto $[0, \bar{c}]$; for an incorrect rollout, it projects $z_i + \min(0, \Delta u_i)$ onto $[\underline{c}, 0]$. These intervals establish sign separation. At fixed z_i, the correct branch responds only to positive structural deviations and the incorrect branch only to negative deviations. Monotone clipping preserves this ordering, with ties at saturation.

For a prompt group consisting of exactly two correct rollouts ($G = 2$), with $r_1 = r_2$ and $u_1 > u_2$, standardization gives $z_1 = z_2 = 0$. Both belong to the same stratum, so $\Delta u_1 = (u_1 - u_2)/2$ and $\Delta u_2 = -(u_1 - u_2)/2$. Substitution into Eq. 4 yields $\widehat{A}_1 = \min\{\bar{c}, (u_1 - u_2)/2\}$ and $\widehat{A}_2 = 0$, proving the tie claim. The corresponding incorrect pair gives $\widehat{A}_1 = 0$ and $\widehat{A}_2 = \max\{\underline{c}, -(u_1 - u_2)/2\}$. For a mixed-correctness pair, both strata are singletons and both structural deviations vanish.

Why scalar ties matter. Let $b_i = \max(r^{(i)}_{\text{base}}, \epsilon)$. Whenever $b_i(1 + u_i) \geq 1$, the hierarchical scalar reward saturates at one. For $b_i > 0$, the threshold is $u_i \geq b_i^{-1} - 1$. Thus, different structural scores can produce identical returns before group standardization. ACE accesses u_i before this clipping and retains its admissible within-stratum deviation. The same mechanism applies to ties arising without saturation. Thus, a structural coefficient can remain nonzero after the group-relative return vanishes.

With $G = 2$, min–max normalization maps an unequal pair of component scores to zero and one, and maps a tied pair to $(1/2, 1/2)$. Consequently, the mixture retains component rankings rather than their original distances. The tie proposition concerns differences in this normalized mixture. Its sign bounds constrain rollout coefficients; shared model parameters, policy-ratio clipping, and other loss terms still determine the combined gradient update.

B.4 TGD: Revision, Acceptance, and Update

At update k, initial traces are sampled from π_{θ_k}. The extractor returns a defect type and a localized step, which enter the revision instruction together with the problem and complete initial trace. A fixed teacher proposes $\widetilde{y}$. Traces without a localized structural defect receive an answer-correction or concise-restatement instruction. The next update uses fresh student traces, allowing revision requests to follow changes in the policy.

Let B denote the configured response budget and $\ell(\widetilde{y})$ the length measured for that budget. The acceptance rule is

$$\begin{aligned} a_r(x, y^\star, \widetilde{y}) = \mathbb{1}[\text{answer}(\widetilde{y}) \text{ is correct}]\mathbb{1}[\text{format}(\widetilde{y}) \text{ is valid}] \\ \cdot \mathbb{1}[q_{\text{topo}}(\widetilde{y}) \geq q_{\text{topo}}(y^\star)]\mathbb{1}[q_{\text{cont}}(\widetilde{y}) \geq q_{\text{cont}}(y^\star)]\mathbb{1}[\ell(\widetilde{y}) \leq B]. \end{aligned} \tag{7}$$

The gate evaluates raw trace scores, before group normalization. Non-decrease preserves aggregate structural quality; it allows ties and does not require every flagged location to change. The teacher and student share architecture and parameter count but have distinct checkpoint roles.

On a teacher-revised prefix, the student conditions on $(x, \widetilde{y}_{<t})$, whereas the teacher also receives $(y^\star, P_r)$. The token loss is

$$\text{KL}(p_t \,\|\, q_t) = \sum_{v \in \mathcal{V}_{\text{tok}}} p_t(v) \log \frac{p_t(v)}{q_t(v)}, \tag{8}$$

where $\mathcal{V}_{\text{tok}}$ is the shared token vocabulary. Gradients pass through p_t only; sampled traces, teacher revisions, acceptance decisions, and q_t remain fixed for that update. Equation 5 averages the gated loss over revision attempts. Rejected attempts contribute zero. The original student trace determines the feedback request, and the accepted teacher revision supplies the token targets.

B.5 Evaluation Protocol

Public mathematics tasks use boxed-answer extraction and symbolic equivalence checking, without a rubric grader.

Pass@5 suite. For Table 1, pass@5 is $N^{-1} \sum_{i=1}^{N} \mathbb{1}[\max_{j \leq 5} c_{ij} = 1]$ over five samples at temperature 0.7 and top-$p = 0.95$. Completion caps are 1,536 tokens for GSM8K and MATH, 2,560 for competition and olympiad sets, and 512 for MMLU and GPQA-D. External training budgets are not matched; these rows contextualize end-to-end checkpoints rather than isolate a training treatment.

Controlled pass@1 suite. The controlled study decodes once per item and retokenizes visible text without special tokens. Table 15 instead varies decoding seeds for fixed policies. The former compares observed checkpoints; the latter measures decoding variation, not training-seed uncertainty. Neither protocol is pooled with the pass@5 ledger.

Optimization-log suite. Figure 4 retains online-batch Stage-II trajectories and endpoint aggregates on GSM8K, MATH-500, AIME 2024, and Omni-MATH. Its task mixture and aggregation follow the optimization logs and are not pooled with either primary table. The extended p@5/maj@5 cells below likewise retain their contemporaneous controlled-suite subsets and harness.

C Extended Results

Table 1 reports the nine-task pass@5 comparison at its recorded precision. Appendix tables below expose the checkpoint-level pass@1 and per-benchmark ablation measurements used for the accuracy–length analysis. Accuracy–length claims use only shared-initialization controls.

For scale context, Qwen2.5-7B TopoPRM (55.3) exceeds the complete DeepSeek-R1-Distill-Llama-8B reference (50.7); Qwen3.5-9B attains the highest reported macro-average in Table 1. These checkpoint comparisons motivate the shared-initialization analysis below, which examines response cost and training components under a common backbone.

C.1 Controlled Pass@1 Checkpoints

Table 6: Controlled pass@1 checkpoint comparison on four mathematics benchmarks. † marks checkpoints evaluated with our prompts and answer extractor. **Avg.** is the macro-average over four tasks.

Method	Params	GSM8K	MATH-500	Olympiad Bench	AIME 2024	Avg.
Complete open checkpoints						
Qwen3.5-4B†	4B	57.7	38.6	19.4	3.3	29.8
Qwen3.5-9B-Instruct†	9B	66.6	41.2	26.1	3.3	34.3
Llama-3.1-8B-Instruct	8B	84.5	52.2	15.0	3.0	38.7
Qwen2.5-7B-Instruct	7B	91.6	**75.5**	25.0	7.0	49.8
Controlled Qwen3.5-9B post-training variants						
Stage-I Base	9B	55.2	19.8	11.2	16.7	25.7
Outcome Only	9B	93.3	50.8	31.3	16.7	48.0
w/o Topology	9B	93.0	51.0	29.8	20.0	48.5
w/o Continuity	9B	51.0	21.6	8.2	6.7	21.9
TopoPRM (Full)	9B	**94.1**	51.0	**32.8**	**30.0**	**52.0**

The lower block supplies Figure 3 and Table 2; its variants share the Stage-I initialization. Full and *w/o ACE* include TGD, while the other controlled variants stop after Stage II. The upper block reports complete open checkpoints, with † marking checkpoints evaluated using our prompts and answer extractor. Accuracy–length claims use only the controlled lower block.

C.2 Complete Per-Benchmark Ablation Ledger

Tables 7–10 report the complete eight-benchmark ablation ledger. The *w/o ACE* checkpoint retains TGD and is the closest control for the estimator; topology and continuity removals stop after Stage II and measure end-to-end pipeline sensitivity.

Table 7: Detailed Qwen3.5-9B ablation ledger for GSM8K and MATH-500. Tok. is mean visible response length; Time is total evaluation wall-clock time in seconds.

	GSM8K ($n = 1{,}319$)						**MATH-500** ($n = 500$)					
Variant	p@1	p@5	maj@5	Tok.	Time	Acc/kTok	p@1	p@5	maj@5	Tok.	Time	Acc/kTok
Base	55.2	87.8	55.7	1,536	45.0	35.9	19.8	39.2	21.4	1,536	18.7	12.9
Outcome Only	93.3	97.7	93.6	1,047	34.2	89.1	50.8	59.0	55.2	1,511	18.9	33.6
w/o Topology	93.0	97.7	93.9	946	25.0	98.3	51.0	59.4	55.6	1,506	16.2	33.9
w/o Continuity	51.0	86.1	52.6	1,528	48.3	33.4	21.6	39.6	23.2	1,536	19.1	14.1
w/o ACE	93.8	98.0	94.6	1,006	28.3	93.2	50.8	59.8	55.2	1,515	18.2	33.5
TopoPRM (Full)	94.1	97.9	94.8	796	19.7	118.2	51.0	82.0	54.6	1,454	16.6	35.1

Table 8: Detailed Qwen3.5-9B ablation ledger for OlympiadBench and Omni-MATH. Metric definitions follow Table 7.

	OlympiadBench ($n = 134$)						**Omni-MATH** ($n = 262$)					
Variant	p@1	p@5	maj@5	Tok.	Time	Acc/kTok	p@1	p@5	maj@5	Tok.	Time	Acc/kTok
Base	11.2	25.4	14.2	2,560	26.5	4.4	16.4	33.2	17.9	2,559	52.0	6.4
Outcome Only	31.3	46.3	37.3	2,043	19.7	15.3	41.6	55.3	48.9	1,845	35.2	22.5
w/o Topology	29.8	50.7	38.1	2,468	11.8	12.1	42.4	54.2	45.8	2,430	17.8	17.4
w/o Continuity	8.2	25.4	9.0	2,560	8.7	3.2	14.9	30.5	17.6	2,560	18.5	5.8
w/o ACE	32.8	47.0	35.8	2,491	9.6	13.2	42.0	53.8	47.3	2,463	19.4	17.0
TopoPRM (Full)	32.8	50.7	38.8	1,768	18.1	18.6	42.8	55.7	47.7	1,592	19.0	26.9

Table 9: Detailed Qwen3.5-9B ablation ledger for AIME 2024 and AIME 2025. Metric definitions follow Table 7.

	AIME 2024 ($n = 30$)						**AIME 2025** ($n = 90$)					
Variant	p@1	p@5	maj@5	Tok.	Time	Acc/kTok	p@1	p@5	maj@5	Tok.	Time	Acc/kTok
Base	16.7	33.3	33.3	4,083	8.5	4.1	2.2	4.4	1.1	2,560	5.9	0.9
Outcome Only	16.7	33.3	20.0	2,546	5.6	6.5	13.3	15.6	8.9	2,547	16.7	5.2
w/o Topology	20.0	40.0	30.0	2,560	2.1	7.8	12.2	16.7	10.0	2,560	6.2	4.8
w/o Continuity	6.7	13.3	3.3	2,560	2.3	2.6	1.1	4.4	0.0	2,560	7.0	0.4
w/o ACE	20.0	30.0	23.3	2,560	2.6	7.8	13.3	17.8	12.2	2,560	7.7	5.2
TopoPRM (Full)	30.0	43.3	33.3	2,431	5.4	12.3	15.6	18.9	11.1	2,457	7.7	6.3

Table 10: Detailed Qwen3.5-9B ablation ledger for CNMO 2024 and MMLU. Metric definitions follow Table 7.

	CNMO 2024 ($n = 30$)						**MMLU** ($n = 500$)					
Variant	p@1	p@5	maj@5	Tok.	Time	Acc/kTok	p@1	p@5	maj@5	Tok.	Time	Acc/kTok
Base	10.0	16.7	3.3	2,560	1.9	3.9	28.6	41.4	26.2	512	55.1	55.9
Outcome Only	16.7	33.3	26.7	2,530	5.5	6.6	63.0	84.0	61.0	512	146.2	123.0
w/o Topology	20.0	40.0	26.7	2,560	2.0	7.8	65.8	85.0	63.0	512	32.3	128.5
w/o Continuity	6.7	20.0	3.3	2,560	2.4	2.6	20.2	34.8	18.6	512	80.4	39.5
w/o ACE	20.0	30.0	26.7	2,560	2.6	7.8	61.2	83.1	61.2	508	58.1	120.3
TopoPRM (Full)	30.0	50.0	30.0	2,530	2.5	11.9	68.2	86.4	68.8	508	48.0	134.2

C.3 Structural Credit and Teacher Revision

Training and evaluation protocol. This supplementary suite uses seed 42 and the frozen extractor. Stage II starts from the Qwen3.5-9B Stage-I adapter and runs 200 updates with two completions per prompt and 16 rollouts per update, giving 3,200 rollout records. The completion cap is 4,096 tokens. It uses the hierarchical reward in Eq. 3 and ACE bounds $(\underline{c}, \bar{c}) = (-1, 1)$. The TGD and generic-revision runs start from the same resulting Stage-II checkpoint. Both use the fixed Stage-I teacher at SFT step 1,734, with the same architecture and vocabulary as the student. Stage III uses 512 revision attempts, 64 updates, an eight-attempt accumulation denominator, a 2,048-token generation cap, and an 8,192-token context limit. Rejected attempts contribute zero loss.

The two revision variants share the question sequence, teacher, structural acceptance gate, and update schedule. Generic revision uses a general revision instruction instead of the defect type and localized step supplied by TGD. Both are evaluated with identical prompts, greedy decoding, a 4,096-token completion cap, and the same answer verifier. Response length is measured by retokenizing decoded text without special tokens. GSM8K, MATH-500, OlympiadBench, and AIME 2024 contain 1,319, 500, 675, and 30 items, respectively. This OlympiadBench set is separate from the 134-item subset in Table 2.

Structural credit under scalar ties. The 3,200 rollouts form 1,600 prompt groups. A group has tied scalar rewards when their absolute difference is below 10^{-7}. ACE credit is nonzero when $|\widehat{A}_i| > 10^{-6}$ for at least one member. Of the 1,071 tied groups, 437 (40.8%) retain nonzero ACE credit. These groups have zero scalar group-relative returns, so their nonzero coefficients come from the structural term in Eq. 4. Here correctness clipping alone projects z_i onto $[0, 1]$ for correct answers and $[-1, 0]$ otherwise. Coefficient differences use the same 10^{-6} tolerance. Across all rollouts, 670 coefficients (20.9%) differ from correctness clipping alone, while another 243 (7.6%) change only through that clipping. Recomputing Eq. 4 from the logged inputs matches the stored coefficients within 4.5×10^{-8}.

For example, at update 177, two correct responses to an orange-bag arithmetic question both receive scalar reward one. Their topology scores are 1 and $10/13$, and both continuity scores are zero. Group normalization gives structural mixtures 0.8 and 0.2, leading to ACE coefficients 0.3 and 0. This pair instantiates Proposition 1 with recorded student outputs.

Teacher-revision comparison. Table 11 reports the teacher, shared initialization, and both revision endpoints on all four tasks. On OlympiadBench, TGD solves 169 problems and generic revision solves 143, a difference of 3.85 percentage points. The difference on MATH-500 is 3.60 points. Their mean response lengths increase by 13.4% and 17.3%, respectively, relative to generic revision. GSM8K accuracy decreases by 0.76 points, and the AIME difference corresponds to two questions. This instruction comparison therefore has a different accuracy–length profile from the historical Full/*w/o ACE* pair in Table 2.

Table 11: Teacher and student endpoints in the supplementary seed-42 suite. Cells report greedy pass@1 (%) / mean visible tokens. Both revision variants share the Stage-II initialization.

Configuration	GSM8K	MATH-500	OlympiadBench	AIME 2024
Stage-I teacher	87.34 / 131.6	59.20 / 360.5	22.22 / 684.3	10.00 / 1047.7
Shared Stage-II student	88.86 / 137.3	60.60 / 385.0	23.11 / 758.5	3.33 / 1211.8
Generic revision	87.49 / 122.4	57.00 / 259.3	21.19 / 529.7	0.00 / 438.8
TGD (Full)	86.73 / 126.2	60.60 / 304.2	25.04 / 600.7	6.67 / 783.0

Paired question analysis. We align outputs by item identifier, question, and gold answer. For correctness indicators c_i^{TGD} and c_i^{gen}, the difference in percentage points is $\Delta = 100N^{-1}\sum_i (c_i^{\mathrm{TGD}} - c_i^{\mathrm{gen}})$. We obtain percentile intervals from 50,000 paired question bootstrap resamples. The implementation samples multinomial counts of the observed differences $\{-1, 0, 1\}$ using a NumPy generator seeded at 925 and advanced across tasks in table order. On OlympiadBench, TGD alone solves 57 questions and generic revision alone solves 31. The resulting interval is $[1.19, 6.67]$ points

(Table 12). The intervals describe question variation for these fixed checkpoints, without adjustment for four-task multiplicity. Training-seed variation is not estimated by this resampling.

Table 12: Paired TGD–generic comparison. Counts are correct questions. Differences and intervals are in percentage points.

Task	N	TGD	Generic	Δ	Paired 95% interval
GSM8K	1319	1144	1154	-0.76	$[-2.27, 0.68]$
MATH-500	500	303	285	$+3.60$	$[0.00, 7.20]$
OlympiadBench	675	169	143	$+3.85$	$[1.19, 6.67]$
AIME 2024	30	2	0	$+6.67$	$[0.00, 16.67]$

Acceptance and textual revision. TGD accepts 336 of 512 attempts (65.6%), compared with 379 (74.0%) for generic revision. Among the accepted TGD outputs, 118 change the student text and 218 repeat it exactly. Five improve continuity, and none increases the raw topology score. Equal-score revisions can pass the gate, so acceptance and structural repair are counted separately. In one accepted revision, the teacher inserts an explicit statement that two sets of bags have the same total weight. The answer remains correct, topology remains one, and continuity rises from zero to 0.2667, with generated length changing from 60 to 79 tokens. The example illustrates localized teacher feedback while the aggregate counts describe its frequency.

D Graph Construction and Edge Audit

D.1 Deterministic Construction

Algorithm 1 summarizes the deterministic rule path. Optional language-model edge refinement is disabled in all reported policy runs.

Algorithm 1 Typed dependency extraction from a reasoning trace

Require: Reasoning text y
Ensure: Support DAG $\mathcal{G}_y$ and typed dependency ledger
Extract the reasoning region and segment it into ordered steps $s_1, \ldots, s_T$
for $j = 1, \ldots, T$ **do**
  Parse normalized expressions, claim keys, variables, and the step type of s_j
  for each earlier step s_i, $i < j$ **do**
    Add typed support $i \to j$ for exact expression or claim reuse
    Add expression-overlap or variable evidence when its guard is satisfied
  end for
  if $j > 1$ and s_j is a derivation/conclusion with no structural predecessor **then**
    Add a guarded local-transition edge from s_{j-1}
  end if
end for
Compute conclusion support and step continuity; retain graph-validity diagnostics
return $\mathcal{G}_y$ and the evidence ledger

Expression, claim, and guarded variable edges receive unit support weight. Local implicit-block and sequential fallback edges contribute 0.5 and 0.3, respectively; incoming support saturates at one. A conclusion with a single fallback edge has support strength 0.3, while a unit-weight edge saturates its score. The score quantifies recovered incoming support, not the fraction of all logically necessary premises: the extractor does not enumerate a complete premise set. Forward construction enforces direction consistency and acyclicity.

Continuity and parser fallback. A parsed step receives $h_i = 1$ when an explicit problem-reference cue or an earlier expression or claim grounds it. If recognizable content has no such cue, $h_i = 0$; if neither an expression nor a claim can be extracted, the parser assigns $h_i = 1$. This last case distinguishes unparsed prose from parsed but ungrounded content. With $\eta = T^{-1} \sum_i h_i$, the

continuity score is $\eta[0.8 + 0.2\mathbb{1}(\eta = 1)]$ for $T > 0$, and zero for $T = 0$. The all-grounded case receives a discrete bonus. Both scoring conventions are applied consistently before reward normalization and TGD acceptance.

D.2 Representation Properties

From score to step location. Every structural contribution resolves to a parsed step, a typed edge with triggering spans, or a continuity decision. This decomposition matters because identical graph-level scores can encode different evidence profiles: one exact expression match and several lower-weight fallback links may produce the same saturated support score while carrying different error risks. The typed ledger exposes that distinction directly, which enables the rule-level false-positive analysis in Section 3.4 without fitting a separate explanation model.

Diagnostic non-redundancy. Topology and continuity induce different equivalence classes over traces. Two responses can have the same local grounding rate while differing in whether their conclusions recover non-local premises; conversely, the same supported conclusion can be reached through locally traceable or discontinuous intermediate steps. Keeping both scores explicit preserves these distinctions for reward analysis and defect-conditioned distillation.

Invariances and failure boundary. The representation is invariant to surface changes preserved by normalization, including whitespace, equation delimiters, and several equivalent expression formats. It is not invariant to arbitrary paraphrase, symbol renaming, or mathematically valid step reordering. Normalization standardizes symbolic cues; the scorer then records which of those cues connect the steps. Continuity can recover local traceability when a non-local edge is missed, while outcome verification remains responsible for mathematically coherent but incorrect traces.

D.3 Audit Sampling and Labels

The audit samples 120 traces, stratified by task (GSM8K or MATH) and response length (short, medium, or long). The typed rule ledger covers all 120; 119 traces receive both model annotations and supply the 1,770-pair contingency table in Table 3. Each evaluation pack contains parsed steps and candidate pairs but hides the extractor's decision and edge type. Qwen3-32B and Qwen2.5-32B independently judge whether an earlier step supplies information used later, after which the complete set is manually checked. Precision, recall, and F1 retain the two model annotations and their consensus/union as explicit diagnostic references; raw agreement and Cohen's κ summarize cross-model agreement.

Manual verification checks the complete pack for annotation and parsing errors. Metrics retain the independent model labels as their references, making disagreements between annotators visible in the contingency table.

D.4 Representative Extraction Failures

Post-hoc rule sensitivity. The typed ledger attributes 145 false positives to variable-reference edges, compared with 34 for expression reference, 11 for expression overlap, 9 for implicit blocks, and 82 for order fallback. This concentration motivates a guarded configuration requiring at least two shared variables for variable-only edges and at least 0.06 token overlap for fallback order edges.

The guarded configuration reduces variable-reference false positives from 145 to 47 and raises overall precision from 0.480 to 0.580 (Table 14). It also changes implicit-block and fallback extraction, making this a multi-rule revision rather than a single-threshold test. This post-hoc analysis uses the same labels for rule design and evaluation. The policy results use the original extractor; evaluating the revised rules requires fresh labels and matched retraining.

Table 13: Representative false-positive edges from the held-out audit. Text is shortened only at sentence boundaries. The examples illustrate three distinct extraction failure modes.

Edge type	Earlier step	Later step	Why the edge is spurious
Variable reference	"For $x = 1$, plug in each integer root to determine a."	"Thus, the possible values of a are $\{-71, -27, -11, 9\}$."	A shared symbol is insufficient to identify which case supplies the conclusion.
Order fallback	"The quarters total $10 \times \$0.25 = \$2.50$."	"The dimes total $12 \times \$0.10 = \$1.20$."	Adjacent calculations are parallel branches, not a support relation.
Expression reference	"One hundred patches cost $100 \times \$1.25 = \$125$."	"Selling 100 patches yields $100 \times \$12 = \$1{,}200$."	Reusing the count 100 does not make the cost calculation a premise of revenue.

Table 14: Post-hoc rule sensitivity on the 120-trace primary audit. The guarded configuration was designed after inspecting these labels and is evaluated on the same sample; the changes are descriptive and are not attributed to the policies in Table 1.

	Released extractor			Guarded extractor			
Evidence type	**TP**	**FP**	**Prec.**	**TP**	**FP**	**Prec.**	**ΔPrec.**
Expression reference	49	34	.590	49	34	.590	.000
Expression overlap	19	11	.633	19	11	.633	.000
Implicit block	15	9	.625	26	13	.667	+.042
Order fallback	127	82	.608	196	122	.616	+.009
Variable reference	49	145	.253	23	47	.329	+.076
Overall	259	281	.480	**313**	**227**	**.580**	**+.100**

Table 15: Matched DR1-Qwen-7B attribution check on a fixed 200-item GSM8K subset. Greedy columns use one decode; repeated columns summarize five stochastic draws from each fixed policy. Standard deviations measure decoding variation, not training-seed variation.

	Single greedy decode		Five stochastic draws		
Reward	**p@1**	**Tokens**	**p@1 mean±sd**	**Tokens mean±sd**	**Structural sources**
Outcome only	75.5	278.9	68.0 ± 1.8	198.1 ± 21.8	none
Outcome + length	76.5	277.5	67.9 ± 2.7	194.4 ± 21.7	none
TopoPRM	77.0	438.3	67.8 ± 1.9	201.3 ± 25.1	topology + continuity

E Robustness and Boundary Tests

The following checks examine decoding stability, score association, cross-backbone behavior, and reranking headroom. Each holds a different part of the evaluation fixed, helping identify where the main accuracy–length pattern persists.

E.1 Matched-Decoding Stability

With greedy decoding, TopoPRM reaches 77.0 pass@1 at 438 tokens, versus 75.5/279 for outcome-only and 76.5/278 for outcome-plus-length. Across five stochastic draws, the fixed policies obtain 68.0 ± 1.8, 67.9 ± 2.7, and 67.8 ± 1.9 pass@1, with overlapping mean lengths. The greedy ordering is not stable to decoding variation. This check uses a different backbone and harness from Table 6 and measures decoding variation with fixed checkpoints.

Table 16: Topology score and correctness on 160 matched GSM8K responses per fixed policy. High and low denote the upper and lower score strata used by the archived analysis.

Policy reward	n	**p@1**	**Mean** q_{topo}	$P(\text{correct} \mid \text{high})$	$P(\text{correct} \mid \text{low})$
Outcome only	160	51.25	.925	55.86	40.82
Outcome + length	160	53.12	.928	57.26	41.86
TopoPRM	160	53.12	.939	57.63	40.48

E.2 Structure–Correctness Association

The complete reward has the highest mean topology score, 0.939. Its high-score stratum exceeds its low-score stratum by 17.15 accuracy points, but outcome-only exhibits a similar 15.04-point gap. The score is associated with correctness in this sample, yet the comparison does not show unique discrimination attributable to topology training.

E.3 Cross-Backbone Check

Table 17: Cross-backbone check on 200 items per benchmark with Llama-3.1-8B-Instruct.

	GSM8K		**MATH-500**	
Policy reward	**p@1**	**Tokens**	**p@1**	**Tokens**
Outcome only	85.0	1,008	46.5	3,823
TopoPRM	85.0	619	42.5	3,427

On Llama-3.1-8B-Instruct, TopoPRM preserves GSM8K pass@1 while reducing mean length from 1,008 to 619 tokens. On MATH-500, length falls from 3,823 to 3,427 tokens while pass@1 changes from 46.5 to 42.5. Response shortening holds on both subsets, with an accuracy cost on MATH-500.

E.4 Full Best-of-8 Ledger

Topology-only selection trails the semantic PRM by 3.75 points on GSM8K and 5.00 points on MATH-500. The hybrid ties the PRM on both tasks; both remain below the pool oracle. The fixed fusion yields no aggregate improvement on these pools. The oracle gap measures remaining selection headroom; item-level disagreements would reveal where the two scorers choose different responses.

Table 18: Diagnostic Best-of-8 selection accuracy (%) on one shared Qwen2.5-Math-7B-Instruct candidate pool (n=80 per benchmark; 2,048-token generation cap). PRM uses Qwen2.5-Math-PRM-7B; Hybrid multiplies normalized PRM by $1 + q_{\mathrm{topo}}$. Oracle reports whether any candidate is correct. This diagnostic is separate from the primary 4,096-token policy comparison.

Benchmark	Pass@1	Maj@8	PRM@8	Topo@8	Hybrid@8	Oracle@8
GSM8K	93.75	95.00	95.00	91.25	95.00	96.25
MATH-500	83.75	85.00	83.75	78.75	83.75	92.50

F Additional Analysis and Interpretation

Connecting the evaluations. Cross-lineage comparisons vary initialization; the controlled suite shares it. The TGD-matched contrast tests ACE within the complete pipeline, connecting the observed accuracy–length trade-off with the saturation mechanism in Proposition 1. The supplementary revision comparison fixes the Stage-II initialization and teacher while changing the revision instruction. The edge audit checks support judgments, and the PRM pool fixes generation. Source removals also change stages or budgets, so isolating their effects requires matched, multi-seed training.

Using the extraction audit. Typed edge records connect scores to specific rules. Variable-reference edges contribute the most recorded false positives, suggesting that repeated symbols often conceal different roles. The guarded revision acts on this diagnosis but is evaluated on the labels used to design it. A held-out audit followed by matched retraining would test whether better extraction improves the learned policy.

Dependency depth and operating regime. The Full/no-topology gap is larger on OlympiadBench and AIME 2024 than on GSM8K and MATH-500. A targeted follow-up would stratify examples by dependency depth to test whether non-local support matters more when several intermediate results must be combined. Stable symbolic anchors favor the present extractor; diagrams, paraphrase, and ambiguous symbols weaken its cues. Correct traces may omit explicit premises, while structurally connected traces can still contain invalid deductions. Answer verification and dependency scoring therefore address different parts of reasoning quality.

G Limitations and Future Directions

Shared challenges and current limitations. Recovering semantically valid dependencies from free-form reasoning remains difficult. Surface consistency does not establish that an inference is valid, and final-answer correctness does not verify every intermediate claim. TopoPRM shares this limitation. Its rule-based extractor can miss paraphrased support or connect unrelated uses of a symbol. Forward construction guarantees a DAG but does not detect circular justification in the original text. The continuity fallback can assign credit to unparsed prose, and saturated support scores can conceal missing premises. ACE constrains credit signs without guaranteeing an accuracy gain, while TGD's non-decrease gate can accept an unchanged trace. Empirically, PRM fusion brings no selection gain on the tested pools, and accuracy gains do not transfer uniformly across backbones or decoding settings. The supplementary mechanism study uses one training seed, and source-removal comparisons also vary training stages or budgets, limiting component-level attribution.

Future research. A next step is to combine learned dependency proposals with symbolic or executable checks, drawing on premise-linked verification (Mukherjee et al., 2025) and studies of reasoning DAGs (Zhong et al., 2026). Calibrated edge confidence could weight structural credit and select teacher interventions. Building on structure-aware distillation (Wen et al., 2026), revision objectives could test whether a flagged relation was repaired while preserving valid premises. Matched teacher-call and token budgets, multiple training seeds, and fresh edge annotations would help separate extraction quality, ACE, and revision effects. Extensions to code, tool use, and multimodal reasoning would require support relations grounded in data flow, tool observations, and visual premises rather than unchanged mathematical extraction rules.