CORD: CONJECTURE-AND-REFUTATION DECODING FOR EVIDENCE-GATED REASONING CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Token-level collaboration enables an LLM to assist an SLM at selected reasoning states, but existing methods often rely on Teacher preference to rank local candidates. Such preference does not establish whether replacing the SLM's current action improves its own subsequent reasoning. We propose **CoRD**(Conjecture-and-Refutation Decoding), which reformulates local collaboration as *intervention verification*. At each triggered state, CoRD forms an incumbent–challenger pair, rolls out short candidate-specific Student futures, and asks an LLM verifier to determine whether the incumbent exposes a reasoning error and the challenger avoids it. An override is made only under this asymmetric evidence, followed by a short verified suffix to stabilize the correction. Across six mathematical reasoning benchmarks and two Teacher scales, CoRD outperforms Teacher takeover and S2T in 11 of 12 settings; with a 32B Teacher it raises Math Avg. from 24.81% to 40.50% over Student-only decoding while keeping generation under Student control. Frozen-state counterfactual analysis further shows that accepted interventions achieve 27.27 percentage points of downstream utility, compared with only 1.20 percentage points for rejected interventions under a 16-token verified suffix. Suffix commitment also eliminates the observed token-only damage. These results support moving token-level collaboration from candidate preference toward evidence-gated correction.

1 INTRODUCTION

Large language models (LLMs) have demonstrated strong capabilities in mathematical and multi-step reasoning (Cobbe et al., 2021b; Wang et al., 2022; Yang et al., 2025). However, relying on large models throughout token-level generation incurs substantial computational cost and inference latency. Small language models (SLMs) provide a more efficient alternative for reasoning generation, but their limited capacity may hinder reliable decision making in complex reasoning processes. Therefore, recent studies have explored selective and collaborative inference between models with different capacities, aiming to combine the efficiency of SLMs with the reasoning capability of larger models (Ong et al., 2025; Fu et al., 2025; Pan et al., 2025; Zheng et al., 2025).

Existing collaborative reasoning methods address different aspects of this challenge. Routing and handoff approaches mainly focus on when to involve a stronger model, while token-level collaboration methods use stronger models to selectively assist local decoding decisions (Ong et al., 2025; Fu et al., 2025; Zheng et al., 2025; Huang et al., 2026; Lyu et al., 2026). Recent future-aware approaches further incorporate downstream signals for candidate evaluation (Li et al., 2026). However, selecting a locally preferred action does not necessarily indicate that the resulting intervention improves the Student's future reasoning trajectory. This motivates us to ask whether a local intervention can be verified by comparing Student-induced future trajectories from a shared pre-divergence context.

This question motivates us to view observed disagreement as an *imperfect temporal marker* of potential reasoning drift rather than the exact onset of such drift. Inspired by cognitive studies of counterfactual reasoning, where humans revisit prior representations, construct alternatives, and evaluate their consequences (Byrne, 2007; Van Hoeck et al., 2015), we introduce a *temporally structured counterfactual window*. Instead of directly replacing the disagreement token, the window reconstructs

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.

the shared pre-divergence context and compares candidate-induced future trajectories from the same reasoning state. This backward reconstruction and forward rollout provide complementary evidence for identifying effective local interventions.

This formulation introduces three key challenges. First, effective reasoning correction requires identifying an appropriate intervention context rather than assuming that the observed disagreement corresponds to the actual origin of the drift. Second, alternative actions must be compared under matched Student-specific future trajectories while maintaining practical online inference cost. Third, a successful intervention should not only identify a better candidate branch but also preserve the verified trajectory after correction to prevent immediate Student drift. Therefore, an effective collaborative framework should reconstruct a shared historical context, obtain informative yet efficient counterfactual futures, and commit validated corrections.

Based on this formulation, we propose **CoRD (Conjecture-and-Refutation Decoding)**, an evidence-gated framework for collaborative reasoning correction. At a triggered disagreement state, CoRD constructs incumbent–challenger comparisons and rewinds to a shared pre-divergence anchor state, from which the Student generates candidate-conditioned trajectories extending through the original divergence into a short verification window. A Teacher verifier evaluates the complete local trajectory segments, and CoRD performs correction only when the incumbent is classified as ERROR while the challenger remains NO_ERROR; otherwise, the current Student trajectory is preserved. When accepted, the verified local continuation window is committed and generation resumes from the updated reasoning state, reducing post-intervention trajectory drift. Through this process, CoRD unifies shared-state reconstruction, counterfactual future evaluation, and asymmetric correction within a local intervention framework.

Our main contributions are summarized as follows:

- We formulate token-level Student–Teacher collaboration as a Student-conditioned counterfactual intervention problem. Unlike methods that rely solely on local candidate preference, CoRD evaluates different intervention candidates through the future trajectories induced under the Student's own reasoning dynamics.
- Inspired by cognitive theories of counterfactual reasoning, we introduce a temporally structured counterfactual window. The proposed window reconstructs the context before the observed disagreement through backward rewinding and reveals the downstream consequences of different interventions through future rollout.
- We propose an asymmetric ERROR-to-NO_ERROR verification rule together with verified trajectory commitment. This mechanism provides a conservative correction strategy that reduces unsupported overrides and post-intervention trajectory drift.
- We evaluate CoRD through end-to-end reasoning performance, frozen-state counterfactual analysis, and controlled ablation studies. Our experiments systematically analyze the effects of rewind, counterfactual rollout, asymmetric verification, verified commitment, and the additional collaboration cost.

2 Related Work

2.1 Adaptive small–large model collaboration.

Small–large collaboration methods reduce inference cost by selectively allocating large-model capacity during reasoning. Sequence-level routing methods typically determine which model should handle a request, while more fine-grained approaches further decide when, where, or for how long a large model should assist a smaller model (Ong et al., 2025; Akhauri et al., 2025; Dong et al., 2026; Zheng et al., 2025; Huang et al., 2026; Lyu et al., 2026; Zhang et al., 2026; Zhao et al., 2026; Zeng et al., 2026). These methods mainly focus on the timing, scope, and duration of large-model assistance. In contrast, CoRD focuses on the correction decision after collaboration has been triggered: when an LLM proposes a local intervention at a disagreement state, should that intervention actually be adopted? Rather than directly accepting the local choice at the current state, CoRD further examines how different local decisions affect the Student's subsequent reasoning before deciding whether to apply a correction.

2.2 TOKEN-LEVEL CANDIDATE SELECTION.

Speculative decoding and candidate-reranking methods typically construct a local set of candidate tokens and then select, verify, or rescore them using model likelihoods or auxiliary signals (Leviathan et al., 2023; Cai et al., 2024; Miao et al., 2024; Liao et al., 2025; Li et al., 2023). In small–large model collaboration, this formulation can further constrain the Teacher's decision to the candidate space supported by the Student. In particular, S2T exploits the local sufficiency of the SLM's top-K candidate space and uses the Teacher to rank candidates within this space (Ye et al., 2026). The resulting decision is typically expressed as an individual candidate score or a relative preference among local alternatives. CoRD further separates candidate proposal from intervention verification: local candidates determine which alternatives are worth comparing, while the reconstructed Student trajectories induced by different candidates provide the context for verification. The LLM verifier evaluates these trajectories for reasoning errors and uses the resulting evidence to decide whether a correction should be applied, rather than intervening directly based on local candidate ranking.

2.3 FUTURE-GUIDED REASONING AND VERIFICATION.

Lookahead, speculative reasoning, and verification-based decoding incorporate downstream or intermediate reasoning signals to improve local decisions (Liao et al., 2025; Pan et al., 2025; Shi et al., 2025). For example, speculative reasoning methods allow smaller models to propose intermediate reasoning steps that are subsequently assessed or corrected by a larger model, while FutureBridge uses an answer-verified future to evaluate candidate compatibility with later reasoning outcomes (Pan et al., 2025; Shi et al., 2025; Li et al., 2026). Verification-based methods similarly use explicit correctness signals to assess generated continuations (Cobbe et al., 2021a). CoRD follows this future-guided perspective but shifts the focus from candidate-quality assessment toward verifying the effectiveness of a local intervention, using counterfactually reconstructed Student trajectories to determine whether a proposed correction is sufficiently supported.

3 PRELIMINARIES AND PROBLEM FORMULATION

We formulate autoregressive language model reasoning as a sequential decision process over a finite vocabulary $\mathcal{V}$. Given a query $x \sim \mathcal{X}$, a smaller Student policy π_S generates a reasoning trajectory $\bar{\tau} = (y_1, \ldots, y_T)$ token by token according to $\pi_S(y_t \mid h_t)$, where $h_t = (x, y_{<t})$ denotes the context history at step t and y_j the recorded token at position j. An auxiliary larger Teacher policy π_T provides token-level guidance through its conditional distribution $\pi_T(\cdot \mid h_t)$. The correctness of a complete trajectory is evaluated by a terminal task reward $r(\bar{\tau}) \in \{0, 1\}$. Throughout, a hat ($\hat{y}_j$, $\hat{h}_j$) marks quantities that CoRD reconstructs in place of the recorded ones (§4.1).

From Local Point-wise Likelihood to Student-Conditioned Utility. Standard collaborative decoding methods trigger an intervention at a step t_d when a divergence metric between π_S and π_T exceeds a predefined threshold, and substitute the Student proposal with the Teacher's local favorite:

$$a_t^T = \arg\max_{a \in \mathcal{V}} \pi_T(a \mid h_t). \tag{1}$$

CoRD restricts this search to the Student-supported candidate set $\mathcal{C}_K(h_t) = \mathrm{TopK}_a\ \pi_S(a \mid h_t)$ (§4.1); we keep the unrestricted form here for exposition. Point-wise Teacher likelihood, however, reflects only the Teacher's preference under the static context h_t. It does not account for whether a_t^T can be successfully resolved under the *Student's own downstream autoregressive dynamics* π_S. To formalize downstream impact, we define the **Student-conditioned future utility** of a candidate action a under context h_t as

$$U_S(h_t, a) \triangleq \mathbb{E}_{\bar{\tau} \sim \pi_S(\cdot \mid h_t, a)}\big[r(\bar{\tau})\big] = P\big(r(\bar{\tau}) = 1 \mid h_t, a\big). \tag{2}$$

We call the action currently scheduled to be emitted at t the *incumbent* a^I and an alternative action proposed for the same position the *challenger* a^C. The ideal intervention gain of a^C over a^I is

$$\Delta_t \triangleq U_S(h_t, a^C) - U_S(h_t, a^I), \tag{3}$$

and the oracle intervention rule is $g_t^\star = \mathbb{I}[\Delta_t > 0]$. Estimating Δ_t directly, however, requires full-length Student rollouts and final-answer evaluation for every candidate, which is computationally prohibitive during online inference. CoRD therefore replaces Δ_t by a finite-window counterpart Δ_F for offline analysis (Eq. (10)) and approximates $g_t^\star$ online with a verifier-based gate (Eq. (**??**)).

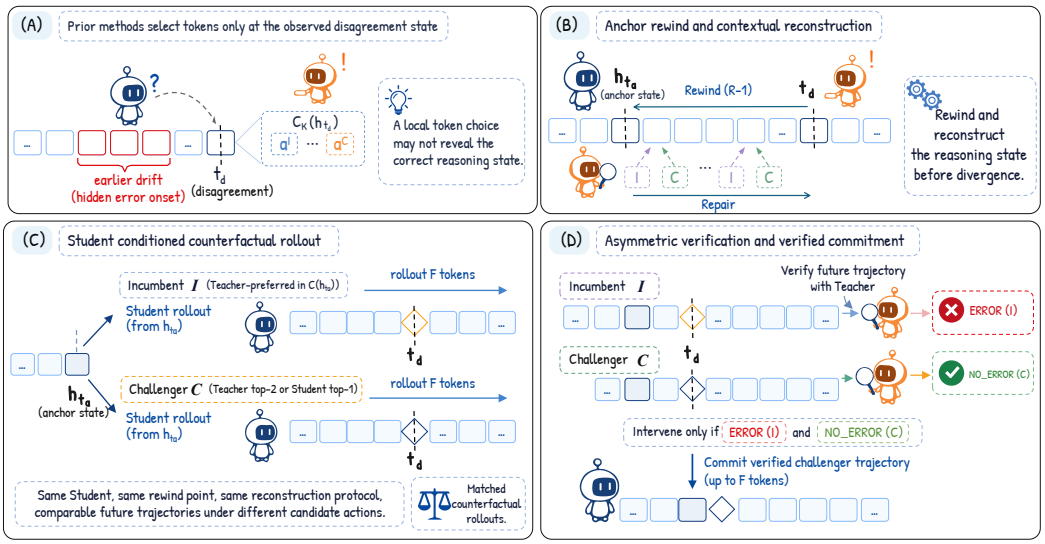

Figure 1: **Schematic of CoRD. (A)** Local token selection at the observed disagreement state t_d may fail to recover the reasoning state before the drift occurred. **(B)** CoRD rewinds to an anchor state h_{t_a} and repairs the pre-disagreement context through token-wise comparisons. **(C)** From the repaired state, CoRD performs matched Student-conditioned counterfactual rollouts for incumbent (I) and challenger (C). **(D)** A Teacher verifier applies the asymmetric rule $\textsc{Error}(I) \wedge \textsc{No_Error}(C)$, and commits the verified challenger trajectory.

The Temporal Lag of Divergence: Why Myopic Interventions Fail. Beyond utility mismatch, a second limitation underlies token-level collaboration. As shown in Figure 1(A), the observable divergence at t_d is typically an *imperfect temporal marker*: the underlying semantic or mathematical drift often begins several tokens earlier, at a position where the two policies still agree on the surface token. An intervention that only replaces y_{t_d} therefore conditions on an already corrupted prefix h_{t_d}, and the correct action at t_d may not exist given that prefix. This motivates inspecting and reconstructing a bounded window of R history tokens before t_d (§4.1) rather than treating t_d as the point of failure.

4 Method: Rewind, Reconstruct, Verify, and Commit (CoRD)

Given a detected Student–Teacher disagreement, CoRD constructs a small set of candidate actions, reconstructs the history immediately before the disagreement, and compares candidate-conditioned Student continuations from the same reconstructed prefix. The Teacher is used to propose and verify local alternatives. A challenger is accepted only when the current incumbent is classified as erroneous and the challenger is not rejected. The procedure consists of trigger and candidate construction and tokenwise historical reconstruction (§4.1), Student-conditioned future rollout (§4.2), and asymmetric verification followed by verified trajectory commitment (§4.3).

4.1 Anchor Rewind and Contextual Reconstruction

Trigger and Candidate Construction. Collaboration is triggered at step t_d when the divergence between π_S and π_T on the shared Student support exceeds a calibrated threshold. At h_{t_d}, CoRD freezes the Student top-K set $\mathcal{C}_K(h_{t_d})$, takes the Teacher-preferred token in this set as the initial incumbent a_0^I, and forms an ordered list of at most $B_d \leq 2$ ratio-gated challengers $a_1^C, \ldots, a_{B_d}^C$: the Teacher top-2 and the Student top-1 candidates whose Teacher probability is at least a fixed fraction of that of a_0^I. If $B_d = 0$, CoRD emits a_0^I and resumes Student decoding; otherwise all candidates remain fixed throughout the steps below.

Anchor State Rewind. In response to the divergence lag diagnosed in Figure 1(A), CoRD rewinds the reasoning sequence by R tokens ($R \geq 1$) to a shared pre-disagreement anchor (Figure 1(B)):

$$t_a \triangleq t_d - R, \qquad h_{t_a} = (x, y_{<t_a}), \qquad \hat{h}_{t_a} \triangleq h_{t_a}, \tag{4}$$

where $\hat{h}_\cdot$ denotes a reconstructed prefix. Anchoring all subsequent evaluations at h_{t_a} ensures that counterfactual branches start from a common state instead of conditioning on the potentially corrupted prefix h_{t_d}.

Tokenwise Historical Reconstruction. For each position $j \in \{t_a, \ldots, t_d - 1\}$, let y_j be the recorded token and $\hat{h}_j$ the prefix after all earlier reconstruction decisions. The historical challenger is the Teacher-preferred token in the Student top-K support at the current reconstructed prefix:

$$c_j = \arg \max_{a \in \mathcal{C}_K(\hat{h}_j)} \pi_T(a \mid \hat{h}_j), \qquad \mathcal{C}_K(\hat{h}_j) = \mathrm{TopK}_a\ \pi_S(a \mid \hat{h}_j). \tag{5}$$

If $c_j = y_j$, the recorded token is retained. Otherwise, two order-swapped Teacher prompts compare y_j and c_j,

$$u_j^{(1)} = P_T(x, \hat{h}_j, y_j, c_j), \qquad u_j^{(2)} = P_T(x, \hat{h}_j, c_j, y_j), \tag{6}$$

each parsed as one of $\{y_j, c_j, \perp\}$. The reconstructed token and prefix are

$$\hat{y}_j = \begin{cases} u_j^{(1)}, & u_j^{(1)} = u_j^{(2)} \in \{y_j, c_j\}, \\ y_j, & \text{otherwise}, \end{cases} \qquad \hat{h}_{j+1} = \hat{h}_j \oplus \hat{y}_j. \tag{7}$$

Thus each historical decision conditions on the corrections already made to its left, and the resulting $\hat{h}_{t_d}$ is shared by every frozen disagreement candidate.

4.2 Student-Conditioned Counterfactual Rollout

Rather than scoring candidates via static probabilities, CoRD projects their downstream viability under the Student's actual dynamics π_S (Figure 1(C)). Let F denote the length of the post-disagreement segment, including the forced candidate as its first token. For every frozen candidate a, CoRD forces a after the reconstructed prefix and generates the remaining $F - 1$ tokens greedily under the Student policy:

$$y_r^a = \arg\max_y \pi_S\left(y \mid \hat{h}_{t_d}, a, y_{<r}^a\right), \qquad r = 1, \ldots, F-1. \tag{8}$$

The complete inspected branch is

$$\tau_F(a) = \hat{y}_{t_a:t_d-1} \oplus a \oplus y_{1:F-1}^a, \tag{9}$$

which contains $R + F$ tokens: R reconstructed historical tokens and an F-token post-disagreement segment. All candidates share $\hat{h}_{t_d}$, π_S, and F, so differences between branches are attributable to the candidate alone.

For frozen-state analysis, let $\bar{\tau}_F(a) = \mathrm{Complete}_{\pi_S}(h_{t_a} \oplus \tau_F(a))$ denote the full Student completion obtained after committing the local branch. The local counterfactual utility and gain are

$$U_{S,F}(h_{t_a}, a) = \mathbb{E}[r(\bar{\tau}_F(a)) \mid h_{t_a}, a], \qquad \Delta_F = U_{S,F}(h_{t_a}, a^C) - U_{S,F}(h_{t_a}, a^I), \tag{10}$$

the finite-window counterparts of U_S and Δ_t in Eqs. (2)–(3). The final-answer reward in Eq. (10) is used only for offline analysis; the online gate uses the local verifier labels below.

4.3 Asymmetric Verification and Verified Suffix Commitment

Asymmetric Intervention Gate. A discrete Teacher verifier $V_T(x, h_{t_a}, \tau) \in \{\text{ERROR}, \text{NO ERROR}, \perp\}$ evaluates the local reasoning validity of each candidate branch in a separate call (Figure 1(D)), where $\perp$ denotes an unparsed or malformed response. For notational convenience, let $v(a) \triangleq V_T(x, h_{t_a}, \tau_F(a))$. The outcome $v(a) = \text{NO ERROR}$ indicates that the verifier identified no explicit reasoning flaw within the local branch; it does not constitute a certificate

of global correctness. For the j-th eligible challenger a_j^C, CoRD applies an asymmetric verification gate against the current incumbent a_{j-1}^I:

$$g_{d,j} = \mathbb{I}\big[V_T(x, h_{t_a}, \tau_F(a_{j-1}^I)) = \text{ERROR} \ \wedge \ V_T(x, h_{t_a}, \tau_F(a_j^C)) = \text{NO ERROR}\big]. \tag{11}$$

Equivalently, $g_{d,j} = \mathbb{I}[v(a_{j-1}^I) = \text{ERROR} \wedge v(a_j^C) = \text{NO ERROR}]$. Unparsed verifier responses fall back to retaining the incumbent. The incumbent is updated sequentially as

$$a_j^I = \begin{cases} a_j^C, & g_{d,j} = 1, \\ a_{j-1}^I, & g_{d,j} = 0, \end{cases} \quad j = 1, \ldots, B_d. \tag{12}$$

Candidate branches are cached within a trigger event, so a branch reused by the two sequential comparisons is neither regenerated nor re-verified.

Proposition 4.1 (Harmful-acceptance bound)**.** *Let H denote a state in which the incumbent is correct and the challenger is incorrect, and let $\alpha_C = \Pr(v(a^C) = \textit{NO ERROR} \mid H)$ and $\epsilon_I = \Pr(v(a^I) = \textit{ERROR} \mid H)$ be the two verifier error rates. The rule that accepts whenever $v(a^C) = \textit{NO ERROR}$ has harmful-acceptance probability α_C, whereas the asymmetric* CoRD *gate satisfies*

$$\Pr(g = 1 \mid H) = \Pr(v(a^I) = \textit{ERROR}, v(a^C) = \textit{NO ERROR} \mid H) \le \min(\epsilon_I, \alpha_C), \tag{13}$$

with equality to $\epsilon_I \alpha_C$ if the two verifier errors are conditionally independent given H.

Proof. The event $\{v(a^I) = \text{ERROR}, v(a^C) = \text{NO ERROR}\}$ is contained in both $\{v(a^I) = \text{ERROR}\}$ and $\{v(a^C) = \text{NO ERROR}\}$, which directly yields the upper bound; conditional independence factorizes its joint probability into $\epsilon_I \alpha_C$. □

The same gate admits a conditional positive-gain interpretation. For a target gain $\gamma > 0$, let $B_\gamma = \{\Delta_F \ge \gamma\}$ with $\pi_\gamma = \Pr(B_\gamma)$, define the miss and false-alarm rates of the gate

$$\beta_F = \Pr(g = 0 \mid B_\gamma), \qquad \alpha_F = \Pr(g = 1 \mid B_\gamma^c), \tag{14}$$

and assume $\Delta_F \ge -M$ on B_γ^c for some $M \ge 0$. When $\Pr(g = 1) > 0$, the accepted interventions satisfy

$$\mathbb{E}[\Delta_F \mid g = 1] \ge \frac{\pi_\gamma(1 - \beta_F)\gamma - (1 - \pi_\gamma)\alpha_F M}{\Pr(g = 1)}, \tag{15}$$

so the expected gain of accepted interventions is positive whenever $\pi_\gamma(1 - \beta_F)\gamma > (1 - \pi_\gamma)\alpha_F M$. This is a conditional safety statement about the gate; it does not guarantee that every accepted challenger is correct.

Verified Trajectory Suffix Commitment. In single-token intervention schemes, control reverts to the unguided Student immediately after modifying y_{t_d}, frequently causing the Student to relapse into the same error pattern in subsequent steps. To lock in verified reasoning progress, CoRD commits the inspected local window instead of replacing only the disagreement token. Let $a^\star = a_{B_d}^I$ be the selected incumbent after all $B_d \ge 1$ eligible comparisons. With commitment cap C,

$$\ell = \min(C, R + F), \qquad h^{\text{new}} \leftarrow h_{t_a} \oplus \text{Prefix}_\ell(\tau_F(a^\star)), \tag{16}$$

where the prefix is truncated further by EOS and the remaining output budget when necessary. The written-back window $y_{t_a:t_a+\ell-1}$ is thus fixed to the selected branch without regeneration, which prevents immediate token-level relapse inside the committed window; it does not establish the final correctness of the selected branch. Ordinary Student decoding then resumes from h^{new}.

5 Experiments

5.1 Experimental Setup

Benchmarks and Metrics. We evaluate CoRD on a diverse set of reasoning benchmarks, including GSM8K (Cobbe et al., 2021b), MATH (Hendrycks et al., 2021), AIME 2023–2025 (AI-MO Team, 2024; Zhang & Math-AI Team, 2024; 2025), and OlympiadBench (He et al., 2024). These benchmarks cover a wide range of mathematical reasoning tasks, from elementary arithmetic problems

Table 1: **End-to-end reasoning performance across collaboration strategies.** We report $\mathrm{MeanAcc}_3$ (%) and macro-average ($\mathrm{Math\ Avg.}$). Main results on mathematical reasoning benchmarks. CoRD achieves competitive and consistent performance across different Teacher scales while preserving Student-controlled decoding. Best collaborative results are in **bold**; CoRD is shaded in light blue.

Teacher	Benchmark	Non-Collaborative		Collaborative (Student + Teacher)		
		Student-only	Teacher-only	Takeover	S2T	**CoRD (Ours)**
32B	GSM8K	68.84	95.84	93.46	92.53	**93.94**
	MATH	51.67	83.69	74.29	77.14	**80.48**
	OlympiadBench	18.35	48.01	40.06	39.60	**44.14**
	AIME 2023	0.00	6.67	6.67	**10.00**	6.67
	AIME 2024	6.67	16.67	6.67	6.67	**11.11**
	AIME 2025	3.33	10.00	3.33	3.33	**6.67**
	Math Avg.	24.81	43.48	37.41	38.21	**40.50**
7B	GSM8K	68.84	90.92	86.53	86.61	**90.02**
	MATH	51.67	78.57	67.14	67.50	**73.69**
	OlympiadBench	18.35	37.46	30.73	30.28	**36.85**
	AIME 2023	0.00	10.00	6.67	6.67	**8.89**
	AIME 2024	6.67	16.67	3.33	3.33	**10.00**
	AIME 2025	3.33	6.67	0.00	0.00	**5.56**
	Math Avg.	24.81	40.05	32.40	32.40	**37.50**

to competition-level problem solving. To ensure statistical robustness and fair comparison under stochastic decoding, we report the mean exact match accuracy across three independent runs with distinct random seeds ($\mathrm{MeanAcc}_3$) alongside the macro-average ($\mathrm{Math\ Avg.}$) across benchmarks for all evaluated methods. Additional efficiency measurements, including Teacher verification calls and extra generation cost, are reported in the Appendix.

Models and Baselines. We use Qwen2.5-1.5B-Instruct as the Student model and Qwen2.5-7B/32B-Instruct as Teacher models (Qwen et al., 2025). All model parameters remain frozen, and collaboration is performed only during inference. We compare CoRD with several baselines, including standalone Student decoding (Student-only), standalone Teacher decoding (Teacher-only), Teacher takeover, and a Student-to-Teacher collaborative baseline, S2T (Ye et al., 2026). For fair comparison, all methods use the same Student and Teacher model configurations, consistent decoding configurations, and the same benchmark evaluation settings.

Collaboration and Implementation Protocol. Unless otherwise specified, CoRD uses the default configuration of $K = 16$, $R = 8$, and $F = 8$, where K denotes the candidate pool size, R denotes the length of the reconstructed rewind segment and the candidate token at the detected disagreement position, and F denotes the future rollout horizon. Candidate construction uses fixed Teacher ambiguity-ratio 0.5 and Student candidate likelihood-ratio thresholds 0.5 for candidate filtering. Additional implementation details, including threshold settings and hyperparameter analyses, are provided in the supplementary material.

5.2 Experimental Results

Main Results: Does CoRD Improve End-to-End Reasoning? Table 1 summarizes the end-to-end reasoning performance under different Teacher–Student collaboration strategies. CoRD achieves strong overall performance across different Teacher scales while preserving Student-controlled decoding. With the 32B Teacher, CoRD improves Math Avg. from 24.81% under Student-only decoding to 40.50%. It also surpasses Teacher takeover (37.41%) and S2T (38.21%) under the same setting. Under the 7B Teacher setting, CoRD achieves a Math Avg. of 37.50%, substantially improving over Student-only decoding while remaining competitive with other collaborative strategies. These results demonstrate that CoRD effectively leverages Teacher feedback through localized trajectory repair and future verification, enabling improved reasoning performance while preserving Student-controlled decoding.

Table 2: Component ablation on MATH.

Method	MeanAcc$_3$
CoRD (Ours)	**80.48**
w/o Tokenwise Historical Reconstruction	78.57
w/o Student-conditioned Counterfactual Rollout	80.20
w/o Asymmetric Intervention Gate	78.57
w/o Verified Trajectory Suffix Commitment	78.10

Table 3: Temporal window size.

R	F	Window	MeanAcc$_3$
4	4	8	80.48
8	8	16	**80.48**
16	16	32	78.81
24	24	48	77.86

Table 4: Temporal budget allocation.

R	F	MeanAcc$_3$
1	15	78.52
4	12	79.64
8	8	**80.48**
12	4	79.76
16	0	78.33

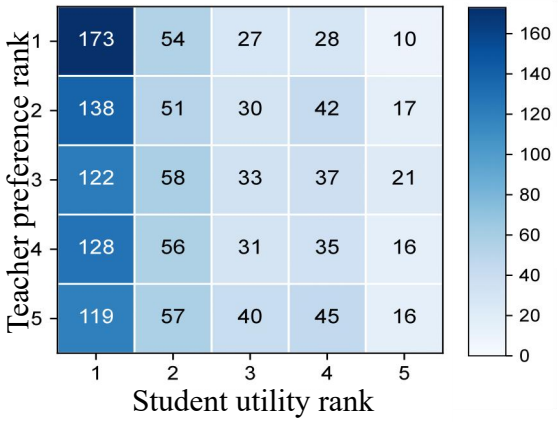

Figure 2: Alignment between Teacher preference and Student-specific utility ranking. Each cell denotes the number of candidate states with the corresponding Teacher rank and Student utility rank.

Consistency across reasoning benchmarks. CoRD maintains stable performance across different Teacher scales and reasoning difficulties. With the 32B Teacher, CoRD achieves competitive results on challenging benchmarks, including 44.14% on OlympiadBench and 6.67% on AIME 2023. When using the weaker 7B Teacher, CoRD remains effective on these challenging reasoning tasks, achieving 36.85% on OlympiadBench and 8.89% on AIME 2023. The consistent performance across Teacher scales indicates that selective trajectory reconstruction and future verification remain effective even when Teacher capability is reduced.

5.3 Why Is Teacher Preference Insufficient?

Figure 2 compares Teacher likelihood ranks with Student-utility ranks over 311 non-degenerate MATH decoding states. For each state, we evaluate eight Student candidates using two candidate-conditioned Student rollouts and define utility by exact-answer success. The ranking distribution is not concentrated near the diagonal, indicating that Teacher preference is not fully aligned with downstream Student utility. Therefore, local Teacher preference alone may be insufficient for intervention decisions, motivating explicit future trajectory evaluation.

5.4 Ablation Studies

We conduct ablation studies to evaluate the contribution of individual components in CoRD. We remove four components respectively: *tokenwise historical reconstruction*, *student-conditioned counterfactual rollout*, *asymmetric intervention gate*, and *verified trajectory suffix commitment*. All variants use the same model configuration and decoding settings. As shown in Table 2, removing historical reconstruction, counterfactual rollout, asymmetric intervention gate, and verified commitment reduces MeanAcc$_3$ from 80.48% to 78.57%, 80.20%, 78.57%, and 78.10%, respectively. These results demonstrate that recovering prior context, evaluating future trajectories, and conservatively committing verified corrections are all important for reliable reasoning correction.

Table 5: Utility enrichment in the frozen candidate comparison cohort.

Dataset	N_A/N_R	Accepted	Rejected	ΔU [95% CI]
MATH	$11/166$	$+27.27$	$+1.20$	$+26.07$ [-1.23,55.85]
GSM8K	$17/272$	$+11.76$	-2.94	$+14.71$ [1.09,30.99]

5.4.1 Temporal Window Analysis

How Does the Temporal Verification Window Affect Performance? We study the effect of temporal verification window size by varying the rewind length and future horizon with balanced allocation. As shown in Table 3, increasing the window size from 8 to 16 tokens maintains the same performance (80.48%), while larger windows of 32 and 48 tokens reduce accuracy to 78.81% and 77.86%, respectively. This suggests that excessively long temporal windows introduce additional verification cost without providing effective benefits. We therefore adopt the 16-token window ($R = 8, F = 8$) as the default configuration, achieving a favorable balance between temporal context recovery, future verification, and inference efficiency.

How Should the Temporal Budget Be Allocated? Given a fixed temporal verification budget, we further investigate how to allocate tokens between historical reconstruction and future trajectory verification. Table 4 compares different R–F allocations under a similar verification budget. The results show that extreme allocations favoring only historical reconstruction or future simulation degrade performance, while balancing the two components yields better intervention decisions. The default setting of $R = 8, F = 8$ achieves the best performance, demonstrating the effectiveness of jointly modeling prior reasoning context and downstream trajectory consequences.

5.5 Mechanism Analysis

Analysis Cohort. We analyze the same frozen $R = 8$, $F = 8$ counterfactual cohort, containing 177 incumbent–challenger comparisons from MATH and 289 from GSM8K. Candidate identities, verifier labels, and outcomes are fixed throughout, enabling verification analysis without rerunning decoding.

Does Verification Identify Higher-Utility Interventions? For each comparison, we compute the utility gain $\Delta U_i = U(\tau_i^C) - U(\tau_i^I)$ between challenger and incumbent trajectories. As shown in Table 5, accepted interventions achieve larger gains than rejected ones: on MATH, $+27.27$ pp versus $+1.20$ pp, and on GSM8K, $+11.76$ pp versus -2.94 pp. These results suggest that future trajectory verification provides useful signals for selecting higher-utility replacements.

6 Conclusion

In this work, we formulate token-level small–large model collaboration as an intervention verification problem rather than candidate selection. CoRD evaluates candidate-induced future trajectories under the Student model and intervenes only when the current trajectory is verified erroneous and an alternative provides a valid improvement. By combining future verification with verified commitment, CoRD enables more reliable local correction while preserving Student-controlled reasoning. Our results show that effective collaboration requires not only finding promising candidates, but also verifying when intervention is beneficial.

AI use statement

In this work, we used generative AI tools for writing and editing parts of the experimental code, for analyzing the experimental data and computing statistics, and for polishing the language of this manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

AI-MO Team. AIMO Validation AIME. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/AI-MO/aimo-validation-aime.

Yash Akhauri, Anthony Fei, Chi-Chih Chang, Ahmed F. AbouElhamayed, Yueying Li, and Mohamed S. Abdelfattah. Splitreason: Learning to offload reasoning. *arXiv preprint arXiv:2504.16379*, 2025. URL https://arxiv.org/abs/2504.16379.

Ruth M. J. Byrne. Précis of the rational imagination: How people create alternatives to reality. *Behavioral and Brain Sciences*, 30:439–440, 2007. doi: 10.1017/S0140525X07002662.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *International Conference on Machine Learning (ICML)*, 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*, 2021a.

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, 2021b. URL https://arxiv.org/abs/2110.14168.

Ziming Dong, Hardik Sharma, Evan O'Toole, Jaya Prakash Champati, and Kui Wu. Pay for hints, not answers: Llm shepherding for cost-efficient inference. *arXiv preprint arXiv:2601.22132*, 2026. URL https://arxiv.org/abs/2601.22132.

Tianyu Fu, Yi Ge, Yichen You, Enshu Liu, Zhihang Yuan, Guohao Dai, Shengen Yan, Huazhong Yang, and Yu Wang. R2R: Efficiently navigating divergent reasoning paths with small-large model token routing. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4135. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b39cef2ef90591cffdc9c674cd55bebe-Abstract-Conference.html.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, 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. *arXiv preprint arXiv:2402.14008*, 2024.

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. In *NeurIPS Datasets and Benchmarks*, 2021.

Chengsong Huang, Tong Zheng, Langlin Huang, Jinyuan Li, Haolin Liu, and Jiaxin Huang. RelayLLM: Efficient reasoning via collaborative decoding. *arXiv preprint arXiv:2601.05167*, 2026.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning (ICML)*, 2023.

Quanquan Li, Hongbo Zhang, Yihe Chi, Jingyu Li, Xidong Xi, Liuyang Song, Hongzhen Zhang, Yuxiang Huang, Jing Ke, Siyuan Ma, Junyi Lin, and Guitao Cao. FutureBridge: Token selection beyond local preference in collaborative decoding. *arXiv preprint arXiv:2608.06819*, 2026.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

Baohao Liao, Yuhui Xu, Hanze Dong, Junnan Li, Christof Monz, Silvio Savarese, Doyen Sahoo, and Caiming Xiong. Reward-guided speculative decoding for efficient LLM reasoning. In *International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 37555–37572, 2025.

Niqi Lyu, Pengtao Shi, Wei Qiu, Jianlin Zhong, Sicong Xia, Jianyao Ma, and Yicheng Ding. PyroDash: Cost-efficient token-level small-large language model collaborative inference. *arXiv preprint arXiv:2607.20327*, 2026.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, et al. Specinfer: Accelerating large language model serving with tree-based speculative inference and verification. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, pp. 932–949, 2024.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, Mohammed Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In *International Conference on Learning Representations*, volume 2025, pp. 34433–34448, 2025.

Rui Pan, Yinwei Dai, Zhihao Zhang, Gabriele Oliaro, Zhihao Jia, and Ravi Netravali. SpecReason: Fast and accurate inference-time compute via speculative reasoning. *arXiv preprint arXiv:2504.07891*, 2025. URL https://arxiv.org/abs/2504.07891.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Junhan Shi, Yijia Zhu, Zhenning Shi, Dan Zhao, Qing Li, and Yong Jiang. SpecCoT: Accelerating chain-of-thought reasoning through speculative exploration. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24405–24415. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1326. URL https://aclanthology.org/2025.findings-emnlp.1326/.

Nicole Van Hoeck, Patrick D Watson, and Aron K Barbey. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in human neuroscience*, 9:420, 2015.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Wenxuan Ye, Yangyang Zhang, Xueli An, Georg Carle, and Yunpu Ma. Select to think: Unlocking slm potential with local sufficiency, 2026. URL https://arxiv.org/abs/2604.26940.

Wenhao Zeng, Xuteng Zhang, Yuling Shi, Chao Hu, Yuting Chen, Beijun Shen, and Xiaodong Gu. GlimpRouter: Efficient collaborative inference by glimpsing one token of thoughts. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17850–17864. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.885. URL https://aclanthology.org/2026.findings-acl.885/.

Chuang Zhang, Zizhen Zhu, Yihao Wei, Bing Tian, Junyi Liu, Henan Wang, Wang Xavier, and Yaxiao Liu. Confidence-calibrated small-large language model collaboration for cost-efficient reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4480–4501. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.208. URL https://aclanthology.org/2026.eacl-long.208/.

Table 6: Default CoRD notation and configuration, following the main-text convention.

Symbol	Meaning	Default
K	Number of Student-proposed candidates ranked by the Teacher	16
R	Maximum number of reconstructed historical tokens	8
F	Post-disagreement segment length, including the forced candidate	8
C	Maximum number of tokens committed from the entire selected window	16

Table 7: Decoding parameters for collaborative trajectories.

Parameter	Value	Parameter	Value
Seed	2025	Complete trials	3
Live temperature	0.7	Live top-p	0.8
Live top-k	20	Local rollout	Greedy
Verifier decoding	Greedy	Verifier output cap	6 tokens
Student support	64	Warm-up steps	3
KL quantile	0.99	Acceptance margin	1

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2024. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/math-ai/aime24.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2025. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/math-ai/aime25.

Yi Zhao, Yajuan Peng, Cam-Tu Nguyen, Zuchao Li, Xiaoliang Wang, Xiaoming Fu, and Hai Zhao. TrigReason: Trigger-based collaboration between small and large reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6700–6718. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.333. URL https://aclanthology.org/2026.findings-acl.333/.

Wenhao Zheng, Yixiao Chen, Weitong Zhang, Souvik Kundu, Yun Li, Zhengzhong Liu, Eric P. Xing, Hongyi Wang, and Huaxiu Yao. CITER: Collaborative inference for efficient large language model decoding with token-level routing. *arXiv preprint arXiv:2502.01976*, 2025.

A Implementation Details

Unless otherwise specified, we use $K = 16$, $R = 8$, $F = 8$, and $C = 16$. Here R counts reconstructed historical tokens and F counts the post-disagreement segment, including the forced candidate as its first token. The default window contains eight historical tokens, the candidate, and seven subsequent Student tokens. The commitment cap C counts the entire selected window, including reconstructed history and the candidate, rather than only the post-candidate suffix.

A.1 Runtime, Decoding, and Prompts

Workers colocate the Student and Teacher on one NVIDIA H200, using local Qwen2.5 Instruct checkpoints with a shared vocabulary, bfloat16 weights, and evaluation mode. The environment uses PyTorch 2.5.1+cu124, Transformers 4.57.6, and Datasets 5.0.0. Evaluation uses no training, retrieval, or external API calls. Reserved supervision tokens are masked in live decoding, candidate construction, and rollouts.

Prompts are zero-shot and use the checkpoint chat template, requesting step-by-step reasoning and a final \boxed{answer}. The future verifier receives the question, visible prefix, candidate, and Student continuation for one branch at a time, and is instructed to return ERROR or NO_ERROR.

Table 8: One tokenwise CoRD intervention cycle.

Stage	Operation	Output
1	Compute KL on the shared Student top-64 support	Trigger decision
2	Rank the Student top-K tokens with the Teacher	Initial incumbent
3	Apply the ordered candidate-ratio gates	Eligible challengers
4	Reconstruct history token by token; force each candidate and generate up to $F-1$ future tokens	Cached branches $\tau_F(a)$
5	Verify each uncached branch independently	Cached labels or $\perp$
6	Apply the asymmetric acceptance rule	Selected branch
7	Commit up to C inspected tokens, subject to EOS and the output budget	Updated live prefix

Table 9: Sequential candidate gates at the original disagreement prefix. The denominator is the frozen initial incumbent a_0^I.

Gate	Challenger	Ratio	Threshold
Teacher ambiguity	Teacher top-2 $a_{T,2}$	$\rho(a_{T,2})$	≥ 0.5
Student disagreement	Student top-1 $a_{S,1}$	$\rho(a_{S,1})$	≥ 0.5

A.2 Hyperparameters and Calibration

After three warm-up steps, both next-token distributions are renormalized on the Student top-64 support. Let $\mathcal{S}_{64}(h)$ denote that support at prefix h. For $a \in \mathcal{S}_{64}(h)$, define

$$\bar{\pi}_M(a \mid h) = \frac{\pi_M(a \mid h)}{\sum_{b \in \mathcal{S}_{64}(h)} \pi_M(b \mid h)}, \qquad M \in \{S, T\}. \tag{17}$$

The restricted distributions are zero outside this support. Collaboration is triggered when

$$D_{\mathrm{KL}}(\bar{\pi}_S \| \bar{\pi}_T) = \sum_{a \in \mathcal{S}_{64}(h)} \bar{\pi}_S(a \mid h) \log \frac{\bar{\pi}_S(a \mid h)}{\bar{\pi}_T(a \mid h)} > \theta_q. \tag{18}$$

The threshold is frozen after calibration. The Teacher ranks only the Student top-K candidates. Teacher-top-2 and Student-top-1 challengers are eligible when their Teacher-probability ratio to the original Teacher-top-1 candidate is at least 0.5. These gates are checked in that order, with the second comparison using the incumbent selected by the first. Candidate futures and verifier labels are cached within an intervention; a reused branch is neither regenerated nor re-verified. The binary verifier processes branches independently and retains the incumbent on parsing failure; its acceptance margin is one. The selected inspected window is committed without regeneration, subject to C, EOS, and the remaining output budget.

Tables 8–10 summarize the intervention stages, candidate eligibility, and verification decision. Historical reconstruction is shared across candidates; it is not repeated for each challenger. If no challenger is eligible, the Teacher-preferred incumbent is emitted without reconstruction or future verification.

The eligibility ratio in Table 9 is $\rho(a) = \pi_T(a \mid h_{t_d})/\pi_T(a_0^I \mid h_{t_d})$. Both Teacher-ranked tokens belong to the frozen Student top-K set. Candidates equal to the initial incumbent and duplicate challengers are omitted. The second comparison uses the incumbent selected by the first, but eligibility ratios are not recomputed after an acceptance or historical correction. Comparisons, trigger events, and problems therefore have different denominators.

Here retention selects the incumbent branch built on the common corrected history; it does not undo historical reconstruction. The selected inspected tokens are committed without post-verification regeneration.

The trigger threshold is the empirical 0.99 quantile on the reserved calibration examples. Recorded 32B thresholds are 2.0214366 for MATH, 2.7907838 for GSM8K, and 2.2216821 for OlympiadBench. AIME imports the GSM8K threshold: 2.7907838 for the 32B Teacher and 2.7243830 for the 7B Teacher. Candidate-ratio thresholds are fixed without tuning on evaluation answers. Table 11 lists the recorded values and imported calibration sources. Non-AIME calibration uses the first 20 loaded examples, which are excluded from evaluation.

Table 10: Asymmetric verification decisions, including parsing failures.

Incumbent label	Challenger label	Decision
ERROR	NO_ERROR	Accept challenger
ERROR	ERROR	Retain incumbent
NO_ERROR	NO_ERROR	Retain incumbent
NO_ERROR	ERROR	Retain incumbent
Unparsed ($\perp$)	Any	Retain incumbent
Any	Unparsed ($\perp$)	Retain incumbent

Table 11: Recorded frozen KL thresholds and calibration sources.

Teacher	Evaluation cohort	Threshold	Calibration source
32B	MATH	2.0214366	MATH
32B	GSM8K	2.7907838	GSM8K
32B	OlympiadBench	2.2216821	OlympiadBench
32B	AIME 2023–2025	2.7907838	GSM8K
7B	AIME 2023–2025	2.7243830	GSM8K

A.3 Trigger Calibration Diagnostics

The empirical diagnostics in this subsection and Sections D–E use the Qwen2.5-1.5B/32B pair and a diagnostic implementation of candidate selection and future verification. These component-level records are reported separately from the final tokenwise method's end-to-end comparisons.

Table 12 compares two recorded KL-trigger settings on the same 1,299-problem GSM8K cohort. Thresholds are frozen after calibration as specified in Section A.2.

The lower-quantile run records approximately 5.1 times the trigger rate. It evaluates several challengers per triggered state, whereas the 0.99-quantile run applies the deployed ratio gates. Because these policies also affect subsequent trajectories, this comparison is descriptive and does not isolate the effect of the quantile on accuracy or compute.

A.4 Model and Dataset Configuration

The model and cohort configurations are summarized in Tables 13 and 14. Model-scale settings use the same token-ID scoring interface. AIME imports the GSM8K threshold for the corresponding Teacher size rather than estimating a threshold on its evaluation questions. The Student-scale row preserves the additional configuration listed in the original experiment plan; no accuracy for that pair is reported in this appendix.

B Evaluation Protocol

The collaborative MATH cohort contains 280 evaluation problems from a 300-example subset of MATH-500; GSM8K and the filtered OlympiadBench pool contain 1,299 and 654 evaluation problems. Each excludes its first 20 loaded examples for calibration. Each AIME year contains all 30 problems and uses no in-domain calibration. Output caps are 1,024 tokens for MATH/GSM8K and 4,096 for OlympiadBench/AIME. Thus, the MATH results describe this fixed 280-problem cohort, not a full MATH-500 evaluation. The underlying benchmarks are MATH, GSM8K, and OlympiadBench (**???**); AIME uses the complete 2023, 2024, and 2025 yearly cohorts.

Dataset-specific answer extraction is followed by symbolic or numeric equality; AIME answers must be integers in $[0, 999]$.

For N problems, let $b_{ir} \in \{0, 1\}$ indicate correctness on problem i in independent decoding run r. The mean single-run accuracy is

$$\mathrm{MeanAcc}_3 = \frac{1}{3N} \sum_{i=1}^{N} \sum_{r=1}^{3} b_{ir}. \tag{19}$$

Table 12: Recorded KL-trigger statistics on GSM8K. Post-trigger policies differ.

Quantile	Threshold	Decoding steps	Triggers	Rate
0.95	0.8783	382,016	14,120	3.70%
0.99	2.7908	381,202	2,764	0.73%

Table 13: Model configurations described in this appendix.

Setting	Student	Teacher
Default	Qwen2.5-1.5B-Instruct	Qwen2.5-32B-Instruct
Teacher scale	Qwen2.5-1.5B-Instruct	Qwen2.5-7B-Instruct
Student scale	Qwen2.5-0.5B-Instruct	Qwen2.5-32B-Instruct

Accuracies are expressed as percentages; diagnostic event counts are reported separately. Thus, averaging the three complete-run accuracies is equivalent to dividing the total number of correct generations by $3N$. This metric is not pass@3, which counts a problem as solved if any of its three generations is correct.

Under the notation used here, the completed tokenwise MATH configurations are $(R, F) = (7, 1)$ and $(11, 5)$. These runs each contain 280 problems with three complete generations per problem. Both enable tokenwise historical reconstruction. Their saved configurations count the candidate within the rewind span and exclude it from the future horizon; the values above convert those settings to the main-text convention without changing the physical windows. Their per-run correct counts are $(215, 211, 200)$ and $(206, 213, 207)$, respectively, yielding $\mathrm{MeanAcc}_3 = 74.52\%$ for each configuration. These end-to-end runs are distinct from the component-level diagnostic cohorts reported below.

C Additional Sensitivity Analysis

The following completed control series characterize candidate construction and future verification on the Qwen2.5-1.5B/32B MATH cohort. They use the diagnostic implementation without tokenwise historical reconstruction and are not matched end-to-end ablations of the final algorithm. The trigger diagnostics are reported in Section A.3.

C.1 Candidate Pool Size Analysis

Table 15 varies the number of Student-proposed candidates in a completed single-pass MATH control series. All candidates remain inside the Student support; the Teacher cannot introduce an out-of-support token.

The measured accuracy changes little from $K = 8$ to $K = 32$, while $K = 4$ is lower on this run. The candidate and verifier counts also remain similar, because the ratio gates prevent most additional candidates from producing a comparison. These are single-run descriptive results; their small differences should not be interpreted as statistically resolved ordering among $K \in \{8, 16, 32\}$.

C.2 Verification Horizon Sensitivity

The horizon control series in Table 16 uses the recorded configuration name H. These are diagnostic continuation budgets, not matched variations of the final tokenwise method. We retain H to distinguish the saved settings from the main-text convention in which F includes the forced candidate.

Increasing H from 16 to 64 multiplies generated future tokens by 4.81 while changing accuracy by only 0.36 percentage points in this single run. This observation supports short local verification windows as a compute-aware default, but it does not by itself prove that longer horizons cannot help on other datasets or under the tokenwise reconstruction rule.

Table 14: Evaluation cohorts and maximum generated-token budgets.

Dataset	Loaded	Calibration	Evaluation	Token cap
MATH	300	20	280	1,024
GSM8K	1,319	20	1,299	1,024
OlympiadBench	674	20	654	4,096
AIME 2023–2025 (each year)	30	0	30	4,096

Table 15: Candidate-pool sensitivity on MATH (pre-tokenwise control series).

K	Accuracy	Comparisons	Verifier calls	Future tokens	Accepted
4	72.50%	181	362	4,757	6
8	75.71%	174	348	4,979	12
16	75.36%	177	354	5,240	11
32	75.36%	182	364	5,432	14

D Mechanism Analysis and Qualitative Examples

Error localization uses a separate replay of 447 eligible accepted comparisons. The verification example comes from the 177 MATH comparisons in the 466-comparison MATH/GSM8K gate-analysis cohort. The two cohorts are not pooled.

D.1 Intervention Statistics

Table 17 summarizes the 466-comparison diagnostic cohort. A comparison is one incumbent–challenger pair; a trigger can yield zero, one, or two comparisons. The committed-token counts describe the saved diagnostic policy, not the final tokenwise commitment policy.

The method accepts only $28/466 = 6.01\%$ of recorded comparisons across these two diagnostics. Teacher calls equal twice the number of comparisons because each branch is checked independently. The low committed-token rate is a direct consequence of sparse triggering, candidate gates, and the asymmetric burden of evidence; it should not be read as verifier accuracy.

D.2 Verification Selectivity and Correction Outcomes

CoRD accepts a challenger only under the directional transition

$$A_i = \mathbb{I}\big[v(\tau_i^I) = \texttt{ERROR} \wedge v(\tau_i^C) = \texttt{NO_ERROR}\big]\,. \tag{20}$$

As shown in Table 18, only 28 of the 466 comparisons satisfy this condition, giving an acceptance rate of 6.01%. By contrast, a naive rule that accepts whenever the challenger is labeled NO_ERROR would accept both ERROR→NO_ERROR and NO_ERROR→NO_ERROR cases, or $408/466 = 87.55\%$. Thus, the asymmetric gate operates as a conservative override mechanism rather than a general candidate reranker.

Accepting every challenger labeled NO_ERROR can damage an already correct incumbent. Table 19 uses frozen counterfactual completions, so each policy is evaluated at the same recorded states. "Naive" accepts whenever the challenger is labeled NO_ERROR, without requiring the incumbent to be labeled ERROR.

Among directional cases, the corresponding correction precision is $5/(5+0) = 100\%$ for CoRD and $22/(22+25) = 46.81\%$ for the naive rule. These figures support conservative selection on the saved states, but the sample of accepted CoRD corrections is small and the replay is not a replacement for a new end-to-end ablation.

D.3 Failure Cases

In the saved error-localization diagnostic, 114 of 447 eligible accepted comparisons were classified as having an error more than 16 historical tokens before the disagreement. Such cases illustrate a

Table 16: Verification-horizon sensitivity on MATH (pre-tokenwise control series).

H	Accuracy	Comparisons	Verifier calls	Future tokens	Accepted
16	75.36%	177	354	5,240	11
32	75.36%	179	358	10,430	11
64	75.71%	221	442	25,203	25

Table 17: Completed intervention accounting at $q = .99$, $K = 16$, and $C = 16$.

Dataset	Problems	Comparisons	Verifier calls	Accepted	Future tok./prob.	Commit tok./prob.
MATH	280	177	354	11	18.71	0.63
GSM8K	1,299	289	578	17	6.36	0.21

coverage limitation: a local reconstruction may modify a downstream symptom while leaving an earlier mistake in the fixed prefix. These are Teacher-localized positions, not human annotations of causal errors.

The same diagnostic localized 267 errors exactly but could not reproduce a local error in 66 cases. The latter may reflect verifier inconsistency, rather than an error-free trajectory. A `NO_ERROR` label therefore means that no explicit error was identified in the inspected window; it does not certify the final answer or rule out mistakes beyond the horizon.

Historical coverage and its limits. Let t_d^i denote the disagreement position and t_e^i the estimated earliest error position, with $\Delta_i = t_d^i - t_e^i$. The denominator contains the 267 exactly localized and 114 right-censored cases, giving 381 cases:

$$\text{Coverage}(R) = \frac{\#\{i : 0 < \Delta_i \leq R\}}{381}. \tag{21}$$

The 66 cases with no reproducible local error are excluded from this denominator. Errors earlier than the 16-token boundary remain in the denominator as right-censored observations.

Figure 3 and Table 20 show that increasing R from 4 to 8 raises coverage from 20.73% to 42.26%. Increasing it from 8 to 12 and from 12 to 16 adds 14.17 and 13.65 percentage points, respectively. Larger windows expose more Teacher-localized errors but require more historical comparisons. This replay measures historical coverage, not whether tokenwise correction causally improves final-answer accuracy, and does not establish $R = 8$ as an empirically optimal window.

D.4 Future Verification Example

In saved MATH state `46:114:0:0c6ce3f25fab5e05`, the problem is to simplify $x(x(1 + x) + 2x) - 3(x^2 - x + 2)$, whose answer is $x^3 + 3x - 6$. The local alternatives are the discourse tokens *Now* and *Next*.

The challenger continuation reaches the correct expression, whereas the incumbent continuation introduces an extra $-x^2$ term. This frozen replay illustrates future verification; it contains no tokenwise reconstruction trace. The displayed futures are excerpts, not complete solutions. The two candidates are nearly interchangeable discourse markers; their mathematical consequences become visible in the subsequent Student text. The frozen comparison changes one recorded state at a time and is not an independent end-to-end accuracy estimate.

E Computational Cost Accounting

E.1 Verification Cost Analysis

This accounting model describes the algorithm, not a fit to the timings below. Let D count trigger events, $J_d \leq R$ the historical positions reconstructed at event d, and $B_d \leq 2$ its eligible challenger comparisons. Set $J_d = 0$ when no challenger is eligible. The historical-position count, branch-

Table 18: Verifier-label transitions in the shared frozen $C = 16$ counterfactual cohort.

Label transition	MATH	GSM8K	Pooled	CoRD decision
ERROR→NO_ERROR	11	17	28	Accept challenger
NO_ERROR→NO_ERROR	146	234	380	Retain incumbent
ERROR→ERROR	14	16	30	Retain incumbent
NO_ERROR→ERROR	6	22	28	Retain incumbent
Total	177	289	466	

Table 19: Frozen-state correction precision for the asymmetric and naive rules.

Dataset	Rule	Accepted	Rescue	Damage	Rescue/accept	False correction
MATH	CoRD	11	3	0	27.27%	0.00%
MATH	Naive	157	10	7	6.37%	4.46%
GSM8K	CoRD	17	2	0	11.76%	0.00%
GSM8K	Naive	251	12	18	4.78%	7.17%
Pooled	CoRD	28	5	0	17.86%	0.00%
Pooled	Naive	408	22	25	5.39%	6.13%

verification count, and generated Student future-token count satisfy

$$N_{\text{hist}} = \sum_{d=1}^{D} J_d, \tag{22}$$

$$N_{\text{verify}} \le 2 \sum_{d=1}^{D} B_d, \tag{23}$$

$$G_S^{\text{future}} \le 2(F-1) \sum_{d=1}^{D} B_d. \tag{24}$$

Without branch-label reuse, each comparison requires two verifications and the verification bound becomes an equality. With event-local caching, let U_d be the number of distinct candidate branches verified at event d. Then $N_{\text{verify}} = \sum_d U_d$, where $U_d = 0$ for $B_d = 0$ and $U_d \le B_d + 1$ otherwise, and $G_S^{\text{future}} \le (F-1) \sum_d U_d$. EOS and the remaining budget can shorten these rollouts. The forced candidate is not a generated future token.

Historical selection requires scoring the Student top-K at each corrected prefix. It adds at most $2N_{\text{hist}}$ pairwise Teacher generations; positions whose recorded token equals the Teacher-preferred alternative require no pairwise generation. A parse failure or disagreement between presentation orders retains the recorded token.

E.2 Latency Analysis

For sequential execution, a coarse accounting approximation is

$$T_{\text{total}} \approx T_{\text{base}} + \sum_{d=1}^{D} \left[J_d c_{\text{hist}} + B_d (2(F-1)c_S + 2c_V) \right]. \tag{25}$$

Here T_{base} includes live decoding, KL checks, and local reranking; c_{hist} includes historical candidate scoring and bidirectional selection, c_S is time per rollout token, and c_V is time per branch verification. This expression ignores candidate-future and verifier-label reuse and early stopping; only the $F - 1$ tokens after the forced candidate require Student rollout generation. Reconstruction is sequential because each correction changes the next prefix. Identical historical candidates require no pairwise Teacher generation, and cached candidate futures are reused.

To make the base term explicit, write

$$T_{\text{base}} \approx T_{\text{live}} + N_{\text{KL}} c_{\text{KL}} + T_{\text{rank}} + T_{\text{overhead}}. \tag{26}$$

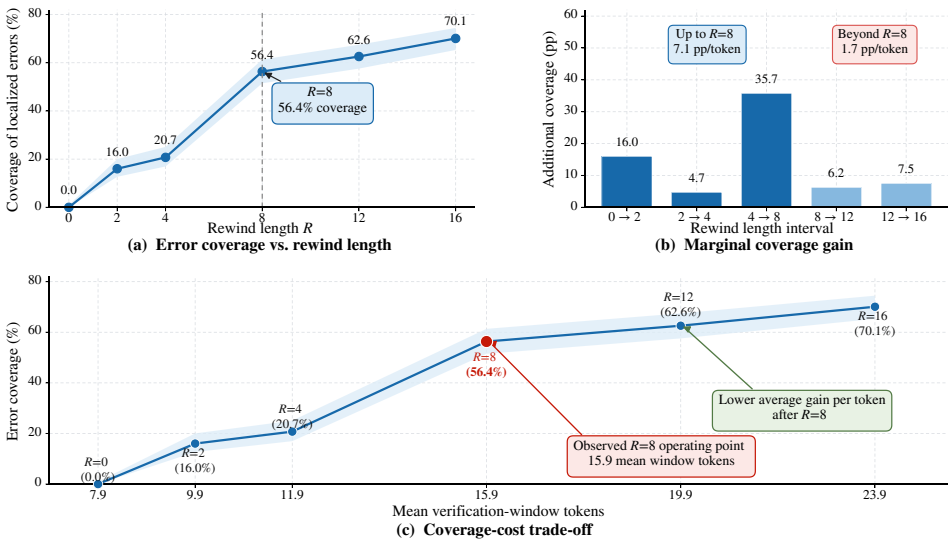

Figure 3: Teacher-localized error coverage (left) and marginal coverage gain (right) in the unified 16-token localization replay. The marked knee is a descriptive geometric statistic, not an end-to-end accuracy optimum.

Table 20: Coverage in the unified Teacher-localization replay.

R	0	2	4	8	12	16
Covered cases	0	61	79	161	215	267
Coverage (%)	0.00	16.01	20.73	42.26	56.43	70.08

Here N_{KL} counts live KL checks, c_{KL} is their incremental cost beyond Student decoding, and T_{rank} is the additional cost of candidate ranking at trigger states. Reused Teacher logits must not be charged again as a separate forward. Historical scoring is included in c_{hist}, not in T_{rank}. T_{overhead} includes bookkeeping and runtime overhead. With caching, the rollout and verification contribution can instead be written as $G_S^{\mathrm{future}} c_S + N_{\mathrm{verify}} c_V$.

Candidate rollouts and branch-verification calls at a fixed corrected prefix are independent and could be batched or parallelized, subject to hardware capacity. This is an execution opportunity, not a measured speedup. The sequential reference execution does not assume such overlap; historical decisions must remain ordered because each changes the next prefix, and challenger acceptance decisions follow the ordered gates.

Recorded diagnostic cost. The table reports verification-call and rollout-token counts alongside end-to-end latency for the diagnostic implementation.

Table 23 reports the corresponding rollout, acceptance, and committed-token counts. These recorded diagnostic counters are distinct from the tokenwise operation bounds above; in particular, their verifier-call counts should not be read as measurements of the event-local label cache in the final specification.

The diagnostic runs generate 18.71 and 6.36 additional Student future tokens per problem, respectively. Their end-to-end timings include runtime overhead and uncontrolled GPU contention, but exclude tokenwise historical correction; they are not latency measurements of the final tokenwise method. The recorded 42.09 and 15.98 seconds per problem were obtained with Student and Teacher colocated on one H200 worker, not in a controlled systems benchmark. No speedup or final tokenwise latency is inferred from these measurements.

Table 21: A future-verification example from the diagnostic MATH replay.

Field	Saved record
Problem	Simplify $x(x(1+x)+2x)-3(x^2-x+2)$.
Gold answer	x^3+3x-6
Local alternatives	Incumbent *Now*; challenger *Next*
Incumbent future	"Now, distribute the -3 outside the parentheses: $[x^2 \ldots$"
Challenger future	"Next, distribute -3 inside the second set of parentheses: $[- \ldots$"
Verifier predictions	Incumbent `ERROR`; challenger `NO_ERROR`
Incumbent outcome	x^3-x^2+3x-6 (incorrect)
Challenger outcome	x^3+3x-6 (correct)

Table 22: Measured cost in the diagnostic implementation ($C = 16$).$Verifiercallsexclude TeacherKL/rerankingforwards$.

Dataset	Problems	Verifier calls	Future tokens	Seconds/problem
MATH	280	354	5,240	42.09
GSM8K	1,299	578	8,267	15.98

F TOKENWISE CORD SPECIFICATION

F.1 TOKENWISE RECONSTRUCTION

At a trigger position t_d, freeze the Student top-K set $\mathcal{C}_d = \mathcal{C}_K(h_{t_d})$, its Teacher-preferred incumbent, and the ordered ratio-gated challengers. With no eligible challenger, emit the incumbent and resume decoding. Otherwise, set $J = \min(R, t_d)$ and $t_a = t_d - J$ for $R \geq 1$, where t_d counts generated tokens before disagreement. This bounds rewind by the prompt boundary; when $t_d \geq R$, it gives the main-text anchor $t_a = t_d - R$. At each historical position $j = t_a, \ldots, t_d - 1$, compare the recorded token y_j with the Teacher-preferred alternative at the already-corrected prefix $\hat{h}_j$:

$$c_j = \arg \max_{a \in \mathcal{C}_K(\hat{h}_j)} \pi_T(a \mid \hat{h}_j), \qquad \mathcal{C}_K(\hat{h}_j) = \mathrm{TopK}_a \ \pi_S(a \mid \hat{h}_j). \tag{27}$$

A pairwise Teacher prompt P_T compares these tokens in both presentation orders at the same already-corrected prefix:

$$u_j^{(1)} = P_T(x, \hat{h}_j, y_j, c_j), \qquad u_j^{(2)} = P_T(x, \hat{h}_j, c_j, y_j). \tag{28}$$

Each parsed response is mapped from its presentation label to the selected token identity in $\{y_j, c_j, \perp\}$. The two orders must agree on a valid token:

$$\hat{y}_j = \begin{cases} u_j^{(1)}, & u_j^{(1)} = u_j^{(2)} \in \{y_j, c_j\}, \\ y_j, & \text{otherwise}, \end{cases} \qquad \hat{h}_{j+1} = \hat{h}_j \oplus \hat{y}_j, \tag{29}$$

Parsing failure or disagreement between presentation orders retains y_j. When $y_j = c_j$, the recorded token is retained without pairwise generation.

For each compared candidate a, append it to the common corrected prefix and generate the remaining $F - 1$ tokens greedily under the Student policy ($F \geq 1$):

$$\tau_F(a) = \hat{y}_{t_a:t_d-1} \oplus a \oplus y^a_{1:F-1}, \qquad y^a_r = \arg\max_y \pi_S(y \mid \hat{h}_{t_d}, a, y^a_{<r}). \tag{30}$$

The generated suffix is empty for $F = 1$; the forced candidate remains in the inspected branch. Verify each branch separately to obtain $v(a) = V_T(x, h_{t_a}, \tau_F(a)) \in \{\texttt{ERROR}, \texttt{NO_ERROR}, \perp\}$, with $\perp$ denoting a parsing failure. A challenger replaces the current incumbent only under

$$g(a^I, a^C) = \mathbb{I}\big[v(a^I) = \texttt{ERROR} \ \wedge \ v(a^C) = \texttt{NO_ERROR}\big]. \tag{31}$$

After all eligible comparisons, the exact selected window (at most $R + F$ tokens) replaces the suffix from t_a, subject to C, EOS, and the remaining output budget. Ordinary Student decoding then resumes without regenerating the inspected tokens.

Table 23: Additional workload counters for the same diagnostic runs ($C = 16$). Committed tokens are those recorded for accepted interventions.

Dataset	Future rollouts	Accepted	Committed tokens
MATH	330	11	176
GSM8K	538	17	272

F.2 Gate Bounds and Their Interpretation

The asymmetric gate requires evidence against the incumbent as well as evidence in favor of the challenger. Let $\mathcal{H}$ be the event that the incumbent is correct and the challenger is incorrect. Define $\epsilon_I = \Pr(v(a^I) = \texttt{ERROR} \mid \mathcal{H})$ and $\alpha_C = \Pr(v(a^C) = \texttt{NO_ERROR} \mid \mathcal{H})$. Then

$$\Pr(g = 1 \mid \mathcal{H}) = \Pr(v(a^I) = \texttt{ERROR}, v(a^C) = \texttt{NO_ERROR} \mid \mathcal{H}) \leq \min(\epsilon_I, \alpha_C). \tag{32}$$

The intersection is contained in each of its two events, which proves the bound without an independence assumption. Conditional independence would give the product $\epsilon_I \alpha_C$, but is not implied by using two calls to the same Teacher.

For the positive-gain statement, let $B_\gamma = \{\Delta_F \geq \gamma\}$, $\pi_\gamma = \Pr(B_\gamma)$, $\beta_F = \Pr(g = 0 \mid B_\gamma)$, and $\alpha_F = \Pr(g = 1 \mid B_\gamma^c)$, for $\gamma > 0$. Assume $\Delta_F \geq -M$ on B_γ^c, with $M \geq 0$. Splitting the accepted gain over these two events gives

$$\mathbb{E}[\Delta_F \mathbb{I}[g = 1]] \geq \gamma \Pr(B_\gamma, g = 1) - M \Pr(B_\gamma^c, g = 1) \tag{33}$$

$$= \pi_\gamma(1 - \beta_F)\gamma - (1 - \pi_\gamma)\alpha_F M. \tag{34}$$

Dividing by $\Pr(g = 1) > 0$ yields the conditional expected-gain bound in the main text. Its positivity depends on the stated rates and gain bounds; neither this argument nor the harmful-acceptance bound certifies individual branches. Final-answer rewards are used only for offline utility analysis, never as inputs to the online gate.