LEARNING WHEN TO IMITATE: COUNTERFACTUAL VALUE MODELING FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) addresses a central mismatch in reasoning distillation by querying the teacher on states actually visited by the student. Yet state alignment does not guarantee useful guidance: in multi-step reasoning, a locally valid teacher step may still lead to a worse final outcome under the student's continuation. Selective variants modulate imitation using uncertainty, disagreement, or teachability, but these proxies do not reveal whether substituting the teacher action improves the current student's eventual outcome. We introduce **Counterfactual On-Policy Distillation (CFOPD)** to estimate this policy-relative intervention value. CFOPD compares teacher and student actions through matched rollouts under the same frozen student policy and amortizes these comparisons with a shared counterfactual value model. The learned values induce two complementary contrasts: action-level contrasts guide supervision between teacher and student alternatives and modulate teacher distillation, while temporal contrasts track progress along the realized student trajectory. Because both depend on the student's continuation ability, the counterfactual supervision is refreshed as the student evolves. Across two Qwen3 teacher–student settings, CFOPD improves Avg@8 over GKD-OPD on all six reasoning benchmarks, with MATH-500 gains from 50.65 to 55.97 and 32.18 to 44.28, respectively. High-budget evaluation further finds the student action preferable at 30–50% of evaluated states despite the teacher's higher overall accuracy, supporting the need for policy-relative action evaluation.

1 INTRODUCTION

Distilling multi-step reasoning from large language models (LLMs) offers a practical route to smaller, more efficient reasoners (Hinton et al., 2015; Kim & Rush, 2016). Autoregressive distillation is complicated by the states that the student creates for itself: because each generated step changes the context for the next, small prediction errors can move the student away from the states represented in fixed teacher demonstrations. On-policy distillation (OPD) addresses this mismatch by training on student-generated trajectories and querying the teacher at the resulting student-visited states (Agarwal et al., 2024). This aligns the training distribution with the contexts the student will encounter at inference time.

State alignment, however, does not establish that the teacher's preferred action is useful at every aligned state. A strong teacher may propose a valid step that leads into a reasoning path the current student cannot successfully continue. The student's own step may be less sophisticated yet place it on a trajectory it can reliably complete. Global teacher superiority therefore does not imply local action dominance. The relevant question is counterfactual: *if the student's action were replaced by the teacher's action, would the same student become more likely to reach a correct solution?*

Selective OPD methods recognize that teacher supervision is heterogeneous and use uncertainty, teacher–student disagreement, or teachability to identify where imitation may be useful (Xu et al., 2026; Wang et al., 2026; Liu et al., 2026; Zhang et al., 2026; Saadi & Wang, 2026). These quantities are informative, but they are observed-path proxies: they do not evaluate what would happen after substituting the teacher's action for the student's. In multi-step reasoning, local confidence or disagreement can differ from downstream utility. Determining whether supervision is beneficial instead requires comparing the two actions at the same state under a shared student continuation policy.

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.

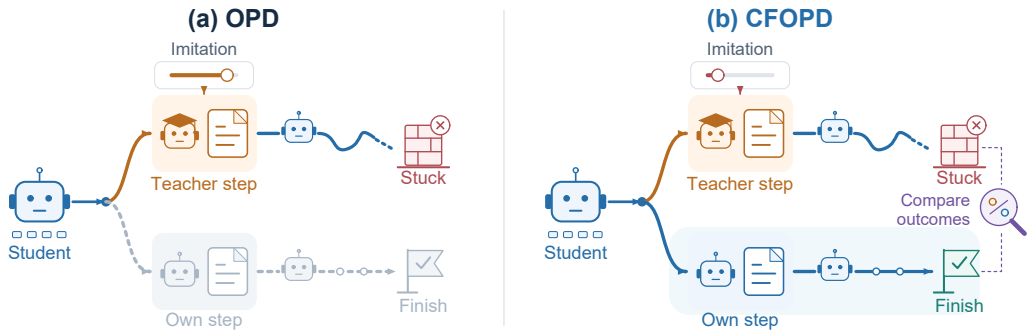

Figure 1: Standard OPD imitates the teacher at student-visited states. CFOPD first evaluates whether a one-step teacher intervention improves the downstream outcome when the same student supplies the continuation.

To address this problem, we introduce **Counterfactual On-Policy Distillation (CFOPD)**, which treats adaptive teacher supervision as policy-relative counterfactual action evaluation. At selected states on an on-policy student trajectory, CFOPD pairs the realized student step with a one-step teacher proposal. The teacher intervenes only at that step; the same frozen student completes both branches. Monte Carlo continuations then estimate the downstream value of each action under the student's continuation policy. Their difference estimates whether replacing the student action with the teacher intervention improves the success probability attainable by the current student.

CFOPD amortizes these rollouts with a shared counterfactual value model. The predicted teacher–student value difference provides an action-level signal for student optimization: it favors locally more useful actions while modulating the strength of teacher distribution transfer. The same model also evaluates the realized student steps; changes in their predicted values provide auxiliary process feedback along the original trajectory. Thus, the counterfactual comparison guides local supervision, while the learned values also indicate *whether the student's reasoning is progressing*. These values are policy-relative because they incorporate the student's ability to continue after an action. As the student improves, an intervention that was previously difficult to exploit may become useful, and the ordering between teacher and student actions may change. CFOPD therefore refreshes its counterfactual supervision as the student evolves.

We evaluate CFOPD on mathematical reasoning under two teacher–student configurations: Qwen3-4B $\rightarrow$ Qwen3-0.6B on MATH and Qwen3-1.7B $\rightarrow$ Qwen3-0.6B on GSM8K (Qwen Team, 2025; Hendrycks et al., 2021; Cobbe et al., 2021). Relative to standard GKD-OPD, CFOPD improves Avg@8 on all six evaluation benchmarks in both settings, including gains of 5.32 and 12.10 points on MATH-500 under MATH and GSM8K training, respectively. The mechanism analysis supports the central premise: approximately 30%–50% of evaluated states favor the student action despite the teacher's stronger overall performance, and the learned model predicts the preferred action with 72.3% sign accuracy on held-out intervention pairs.

Our contributions are threefold:

- We formalize selective OPD through *policy-relative intervention value*, which compares the downstream outcomes of teacher and student actions under the same student continuation policy.
- We introduce **CFOPD**, which amortizes matched counterfactual rollouts with a learned value model and uses the predicted intervention advantage to guide student optimization and adapt teacher supervision on student-generated trajectories.
- Across complementary mathematical reasoning settings, we show consistent gains over standard OPD and provide direct evidence that a globally stronger teacher is not locally preferable at many student-visited states.

2 Related Work

Knowledge Distillation and On-Policy Distillation. Knowledge distillation transfers teacher behavior through token-level distribution matching (Hinton et al., 2015), sequence-level distillation (Kim & Rush, 2016), and rationale supervision (Hsieh et al., 2023). For autoregressive models, training contexts can differ from those induced by the student at inference time. MiniLLM addresses this mismatch with a generation-oriented reverse-KL objective (Gu et al., 2024), while Generalized Knowledge Distillation (GKD) queries the teacher on student-generated trajectories (Agarwal et al., 2024). CFOPD addresses a complementary limitation: even at student-visited states, the teacher's preferred action may not be beneficial under the current student's continuation policy.

Selective and Reliability-Aware Distillation. Selective OPD challenges uniform teacher supervision. TIP identifies informative tokens through student uncertainty and teacher–student divergence (Xu et al., 2026), while Teachability-Aware OPD considers whether supervision lies within the student's effective support (Wang et al., 2026). Other methods assess teacher reliability through trajectory position, prompt-level probes, or local validity (Liu et al., 2026; Zhang et al., 2026; Saadi & Wang, 2026); CROP uses counterfactual input perturbations to identify task-relevant positions (Li et al., 2026). CFOPD instead compares student and teacher actions at the same state under a shared student continuation policy. This matched comparison estimates their relative downstream utility and provides a policy-relative signal for adapting student optimization and teacher supervision.

Process Supervision and Value-Based Reasoning. Process supervision provides intermediate feedback beyond final outcomes (Lightman et al., 2024; Wang et al., 2024a). MiPS, OmegaPRM, and ReST-MCTS* automate such feedback using sampled continuations or search (Wang et al., 2024b; Luo et al., 2024; Zhang et al., 2024). Rewarding Progress and Process Q-value Models further interpret process supervision through value estimation or relative quality modeling (Setlur et al., 2025; Li & Li, 2025). CFOPD extends this perspective from single-trajectory evaluation to counterfactual teacher–student action comparison, using one value model to provide process feedback and policy-relative guidance for student optimization.

3 Method

3.1 Overview

We study on-policy distillation from a teacher π_T to a reasoning student π_θ. Visiting the student's states resolves distribution mismatch, but leaves a second question unanswered: whether the teacher's preferred step is useful to the student that must continue from it. CFOPD makes this question explicit through a *policy-relative action value*: the downstream value of a candidate reasoning step when the current student supplies the remainder of the solution.

CFOPD uses one value function for two comparisons: a *temporal contrast* tracks progress along the realized trajectory, while an *action contrast* directly aligns teacher–student actions and controls teacher distillation.

At round k, the student generates $\mathcal{D}_k$. Selected states and frozen-student counterfactual continuations train CF-PRM, which is then frozen while the same batch is replayed for process feedback, preference learning, and gated OPD (Figure 2).

3.2 Identifying Policy-Relative Intervention Value

Given a reasoning problem x, the current student generates

$$\tau^S = (a_1^S, \ldots, a_L^S) \sim \pi_{\theta_k}(\cdot \mid x), \qquad s_t = (x, a_1^S, \ldots, a_{t-1}^S). \tag{1}$$

where each a_t^S is a reasoning step and s_t is its student-generated prefix. We store the exact tokens of τ^S in $\mathcal{D}_k$ and freeze a snapshot $\bar{\pi}_k = \pi_{\theta_k}$ for counterfactual construction. At a set of selected steps $\mathcal{I}(\tau^S) = \{t_1 < \cdots < t_m\}$, the realized student step a_t^S is paired with a one-step teacher proposal

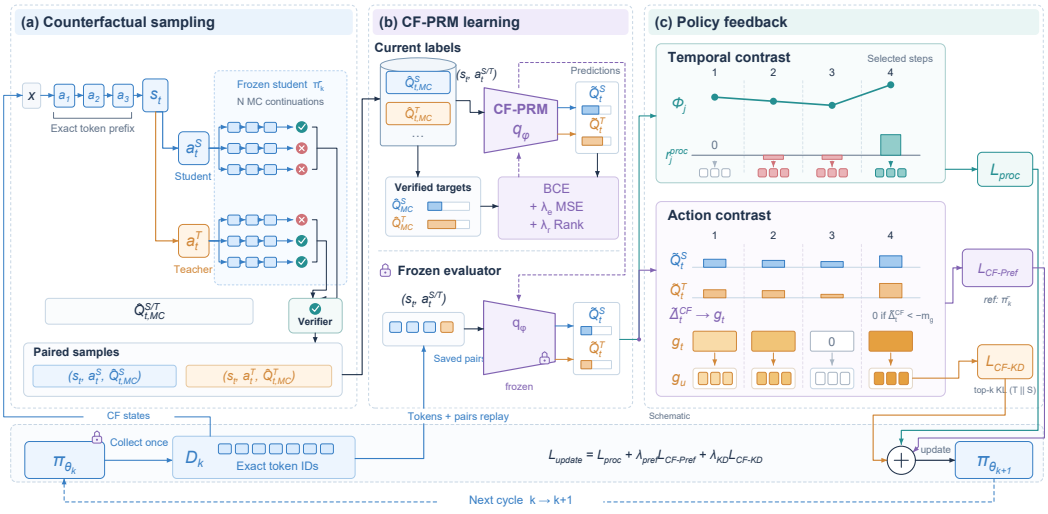

Figure 2: Overview of CFOPD. A student batch is collected once and reused throughout the round. Counterfactual continuations train CF-PRM, after which the evaluated action pairs and original tokens are replayed for preference alignment, process feedback, and counterfactually gated OPD.

$$a_t^T = \mathrm{Decode}\left(\pi_T(\cdot \mid s_t)\right). \tag{2}$$

The teacher intervenes only at the selected step; the frozen student $\bar{\pi}_k$ completes both branches. Holding the prefix, continuation policy, and outcome rule fixed makes the two branches differ only in the local action. For either candidate a, define its policy-relative downstream value as

$$Q_{\bar{\pi}_k}(s_t, a) = \mathbb{E}_{c \sim \bar{\pi}_k(\cdot \mid s_t, a)}\left[V(s_t \oplus a \oplus c)\right], \tag{3}$$

where $V \in \{0, 1\}$ verifies the completed response. This value measures the probability of success that the current student can realize after taking a; it evaluates an action together with the student's ability to continue from that action.

For branch $b \in \{S, T\}$, we estimate the value with N independent frozen-student continuations:

$$\begin{aligned} c_{t,n}^b &\sim \bar{\pi}_k(\cdot \mid s_t, a_t^b), \qquad n = 1, \ldots, N, \\ \widehat{Q}_{t,\mathrm{MC}}^b &= \frac{1}{N} \sum_{n=1}^{N} V\left(s_t \oplus a_t^b \oplus c_{t,n}^b\right). \end{aligned} \tag{4}$$

The paired comparison defines the policy-relative teacher advantage

$$\widehat{\Delta}_{t,\mathrm{MC}}^{\mathrm{CF}} = \widehat{Q}_{t,\mathrm{MC}}^T - \widehat{Q}_{t,\mathrm{MC}}^S. \tag{5}$$

A positive value means that replacing the student step with the teacher proposal increases the success probability attainable by the current student; a negative value favors retaining the realized student step. This matched comparison separates teacher disagreement from teacher advantage. The paired verifier outcomes provide Monte Carlo supervision for CF-PRM.

3.3 Amortizing Counterfactual Evaluation

The Monte Carlo comparison above is used to construct supervision rather than queried repeatedly during student optimization. We amortize it with CF-PRM, $q_\phi(s, a) \in (0, 1)$, which approximates

the policy-relative downstream success probability of an action. Its first-round backbone is initialized directly from the initial student that generated $\mathcal{D}_0$, with a scalar prediction head attached; later rounds warm-start the preceding CF-PRM. No preliminary OPD bootstrap is required.

For $b \in \{S, T\}$, let

$$\widetilde{Q}_t^b = q_\phi(s_t, a_t^b), \qquad \widetilde{\Delta}_t^{\mathrm{CF}} = \widetilde{Q}_t^T - \widetilde{Q}_t^S. \tag{6}$$

Accurate action comparison depends on both calibrated branch values and their relative ordering. We therefore train CF-PRM with

$$\begin{aligned} \mathcal{L}_{\mathrm{value}} &= \sum_{b \in \{S,T\}} \mathrm{BCE}\left(\widetilde{Q}_t^b, \widehat{Q}_{t,\mathrm{MC}}^b\right), \\ \mathcal{L}_{\mathrm{effect}} &= \left(\widetilde{\Delta}_t^{\mathrm{CF}} - \widehat{\Delta}_{t,\mathrm{MC}}^{\mathrm{CF}}\right)^2, \\ \mathcal{L}_{\mathrm{rank}} &= \mathbb{I}\left[\widehat{\Delta}_{t,\mathrm{MC}}^{\mathrm{CF}} \neq 0\right] \left[-\operatorname{sgn}\left(\widehat{\Delta}_{t,\mathrm{MC}}^{\mathrm{CF}}\right) \widetilde{\Delta}_t^{\mathrm{CF}}\right]_+ . \end{aligned} \tag{7}$$

The complete CF-PRM objective is

$$\mathcal{L}_{\mathrm{CF}} = \mathcal{L}_{\mathrm{value}} + \lambda_{\mathrm{eff}} \mathcal{L}_{\mathrm{effect}} + \lambda_{\mathrm{rank}} \mathcal{L}_{\mathrm{rank}}. \tag{8}$$

The value term calibrates each branch, the effect term fits the magnitude of the intervention effect, and the ranking term penalizes an incorrect teacher–student ordering. CF-PRM is then frozen for the corresponding student update, so student-loss gradients do not update ϕ.

3.4 Student Training with Counterfactual Feedback

The student replays the saved trajectories and evaluated action pairs from $\mathcal{D}_k$. Querying the frozen CF-PRM converts the same value surface into complementary temporal and action-level feedback.

Temporal contrast: is the student making progress? For selected steps $t_1 < \cdots < t_m$, let $\Phi_j = q_\phi(s_{t_j}, a_{t_j}^S)$. We use successive value differences as process feedback:

$$r_1^{\mathrm{proc}} = 0, \qquad r_j^{\mathrm{proc}} = \alpha \ \mathrm{clip}\left(\Phi_j - \Phi_{j-1}, -c, c\right), \quad j = 2, \ldots, m. \tag{9}$$

We broadcast the detached r_j^{proc} to tokens $u \in \mathcal{U}_j$ of step $a_{t_j}^S$, with zero weight elsewhere. The process-weighted replay loss is

$$\mathcal{L}_{\mathrm{proc}} = -\frac{1}{\sum_u M_u} \sum_u M_u r_u^{\mathrm{proc}} \log \pi_\theta(y_u \mid h_u), \tag{10}$$

where M_u is the response-token mask.

Action contrast: preference and teacher trust. At the same state, $\widetilde{\Delta}_t^{\mathrm{CF}} = q_\phi(s_t, a_t^T) - q_\phi(s_t, a_t^S)$ determines which evaluated action is preferred. Let $y_t = \operatorname{sgn}(\widetilde{\Delta}_t^{\mathrm{CF}})$, $w_t = |\widetilde{\Delta}_t^{\mathrm{CF}}|$, and $\rho_\theta(a|s) = |a|^{-1} \sum_v \log \frac{\pi_\theta(a_v|s, a_{<v})}{\pi_k(a_v|s, a_{<v})}$, with these value-derived signals detached. We optimize

$$\mathcal{L}_{\text{CF-Pref}} = -\frac{1}{|\mathcal{I}|} \sum_{t \in \mathcal{I}} w_t \log \sigma\left(\beta y_t \left[\rho_\theta(a_t^T \mid s_t) - \rho_\theta(a_t^S \mid s_t)\right]\right). \tag{11}$$

Thus the same teacher–student pair used for counterfactual evaluation is optimized directly, with the frozen round-start policy as reference.

The same contrast also gates distributional distillation. Let $\tilde{g}_t = \sigma((\widetilde{\Delta}_t^{\mathrm{CF}} - \tau_g)/T_g)$ and define

$$g_t = \begin{cases} 0, & \widetilde{\Delta}_t^{\mathrm{CF}} < -m_g, \\ g_{\min} + (g_{\max} - g_{\min})\tilde{g}_t, & \text{otherwise.} \end{cases} \tag{12}$$

Here m_g suppresses teacher actions predicted worse than the student's, τ_g and T_g control the soft gate, and $[g_{\min}, g_{\max}]$ bounds its nonzero weight. Unlike the pairwise objective above, the gated OPD term transfers the full teacher distribution evaluated on the saved student trajectory. We broadcast g_t to the tokens of the corresponding selected step and set $g_u = 1$ outside selected steps. For the top-K forward-KL loss ℓ_u^{KD},

$$\mathcal{L}_{\text{CF-KD}} = \frac{\sum_u M_u g_u \ell_u^{\mathrm{KD}}}{\sum_u M_u}. \tag{13}$$

The two action-level uses are complementary: preference learning directly optimizes the evaluated action pair, whereas the gate treats its predicted advantage as a local trust signal for broader teacher distribution transfer.

Unified student objective. The three signals are combined in one student update:

$$\mathcal{L}_{\mathrm{update}} = \mathcal{L}_{\mathrm{proc}} + \lambda_{\mathrm{pref}}\mathcal{L}_{\text{CF-Pref}} + \lambda_{\mathrm{KD}}\mathcal{L}_{\text{CF-KD}}. \tag{14}$$

The process term evaluates progress along the realized trajectory, the preference term directly aligns the evaluated teacher–student actions, and the gated KD term transfers teacher distributional knowledge only according to its estimated local utility. All three use the same frozen CF-PRM.

3.5 Same-Trajectory Optimization and Refresh

Training starts directly from π_{θ_0}. Within round k, the student collects $\mathcal{D}_k$ once; counterfactual augmentations of its selected states train CF-PRM, after which the saved trajectories and action pairs are replayed for Equation 14. No trajectory batch is inserted between the two updates. In the next round, $\pi_{\theta_{k+1}}$ generates a new batch and refreshes the counterfactual targets. Refreshing is necessary because an action's value is relative to the student's evolving ability to continue from it.

At inference time, only the final student is retained. The teacher, CF-PRM, outcome checker, and counterfactual continuations are training-time components and introduce no deployment cost.

4 Experiments

We evaluate CFOPD from three complementary perspectives: reasoning performance, component contribution, and the validity of the counterfactual mechanism. We study whether CFOPD consistently improves over standard on-policy distillation, how its components contribute to performance, and whether the counterfactual signal provides reliable policy-relative guidance for student optimization.

4.1 Experimental Setup

Models and training settings. We use two Qwen3 teacher–student pairs (Qwen Team, 2025): Qwen3-4B → Qwen3-0.6B for MATH and Qwen3-1.7B → Qwen3-0.6B for GSM8K. Within each setting, methods share the student initialization, prompts, preprocessing, verifier, decoding configuration, and student-update budget.

Training and evaluation data. Methods are trained on the MATH (Hendrycks et al., 2021) or GSM8K (Cobbe et al., 2021) training set, respectively. Both settings are evaluated on MATH-500 (Lightman et al., 2024), AMC23, Minerva Math (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), AIME24, and AIME25. Comparisons are made within each training setting; dataset details appear in Appendix A.

Table 1: Main results under the MATH- and GSM8K-trained settings. Best results are marked in **bold**.

	MATH-500		AMC23		Minerva Math		OlympiadBench		AIME24		AIME25	
Method	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8
MATH-trained	Qwen3-4B → Qwen3-0.6B											
CoT-SFT	47.70	74.00	21.88	62.50	10.98	28.31	20.41	42.35	2.08	6.67	4.16	16.67
Supervised KD	50.00	76.60	27.81	62.50	13.83	34.19	22.51	46.95	3.33	10.00	4.16	10.00
PG-OPD	47.65	72.20	25.62	62.50	12.22	26.47	21.38	42.20	2.50	6.67	2.92	16.67
GKD-OPD	50.65	75.20	25.00	65.00	13.83	30.51	24.88	47.54	2.50	10.00	3.75	13.33
EOPD	52.70	77.20	30.31	65.00	14.39	33.82	26.54	50.52	2.91	10.00	**7.08**	13.33
CFOPD	**55.97**	**81.40**	**33.12**	**67.50**	**15.16**	**35.66**	**27.82**	**52.60**	**3.75**	**13.33**	6.66	**20.00**
GSM8K-trained	Qwen3-1.7B → Qwen3-0.6B											
CoT-SFT	33.53	59.80	10.31	42.50	9.42	23.53	11.85	31.35	0.42	3.33	0.83	6.67
Supervised KD	33.43	61.40	13.13	42.50	10.85	24.26	12.67	33.28	1.25	6.67	0.42	3.33
PG-OPD	28.05	51.20	9.69	35.00	8.46	20.59	9.53	24.81	0.83	6.67	0.83	6.67
GKD-OPD	32.18	60.40	15.63	42.50	9.51	25.37	12.22	31.20	0.83	6.67	0.42	3.33
EOPD	42.70	70.40	23.13	55.00	13.60	30.15	18.48	39.52	2.08	6.67	0.83	6.67
CFOPD	**44.28**	**71.20**	**23.75**	**60.00**	**13.74**	**32.35**	**18.57**	**39.82**	**2.50**	**10.00**	**2.08**	**10.00**

Evaluation protocol. We sample eight responses per problem using the same decoding settings for all methods. **Avg@8** is their mean correctness, while **Pass@8** is the fraction of problems with at least one correct response (Chen et al., 2021).

Baselines. **Offline CoT SFT** trains on fixed teacher reasoning trajectories (Kim & Rush, 2016; Hsieh et al., 2023); **Supervised KD** matches teacher distributions on those trajectories with top-K forward KL (Hinton et al., 2015). **PG-OPD** uses policy-gradient optimization with a reverse-KL-based token signal (Gu et al., 2024). **GKD-OPD**, our primary baseline, applies top-K forward KL on student trajectories (Agarwal et al., 2024). **EOPD** uses entropy-aware on-policy supervision (Jin et al., 2026).

Implementation details. Student rollouts use temperature 0.7 and maximum lengths of $2{,}048$ tokens for GSM8K and $4{,}096$ for MATH. Batch and mini-batch sizes are both 16; student and CF-PRM learning rates are 2×10^{-6} and 1×10^{-5}, respectively. We use eight continuations per branch by default. Further details are in Appendix B.

4.2 Main Results

Table 1 summarizes the main results across the two training settings. CFOPD consistently improves over standard on-policy distillation. In the MATH-trained setting, CFOPD outperforms GKD-OPD on Avg@8 across all six evaluation benchmarks and improves Pass@8 on all six benchmarks. The same pattern is observed under the GSM8K-trained setting, where CFOPD again improves Avg@8 on all six benchmarks and Pass@8 on all six. CFOPD also remains competitive with offline and adaptive distillation baselines across the two settings. These results support the central motivation of CFOPD: training on student-induced states alone does not ensure that every teacher intervention is beneficial. Comparing student and teacher actions under the same continuation policy provides policy-relative guidance for adapting student optimization to local intervention utility. We next examine the contribution of individual components and the learned counterfactual signal.

4.3 Component Ablations

We examine key sources of counterfactual feedback in CFOPD on MATH-500. As shown in Table 2, removing either the process reward or counterfactual guidance degrades performance, showing that trajectory-level feedback and policy-relative action guidance provide complementary benefits. The larger drop without CF guidance highlights the importance of using intervention utility to regulate both student optimization and teacher transfer. Moreover, replacing the student–teacher comparison with a signal based only on predicted teacher value also leads to substantial degradation, showing that teacher quality alone is insufficient: useful supervision depends on the teacher

Table 2: Component ablation results of CFOPD on MATH-500, evaluating the contribution of major component.

Method	Avg@8	Pass@8
CFOPD	**55.97**	**81.40**
w/o Process Reward	53.48	74.60
w/o CF Guidance	51.15	72.20
w/o CF Comparison	51.39	74.40

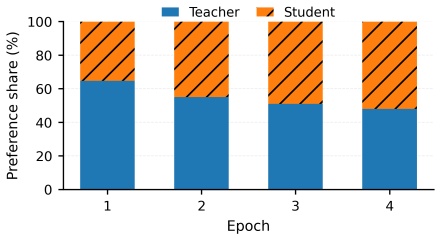

Figure 3: Teacher–student preferences under high-budget counterfactual evaluation.

action relative to the student's own action at the same state. These results support policy-relative comparison, rather than teacher quality alone, as the basis for counterfactual guidance.

4.4 Additional Analyses

The previous experiments establish the overall effectiveness of CFOPD. We now provide additional analyses to better understand its counterfactual supervision signals and training behavior. Additional analyses and experimental details are provided in Appendix B.

Counterfactual utility is more informative than heuristic supervision signals. We compare counterfactual guidance with common heuristic signals at approximately matched activation rates. **Random Matched** selects supervised positions randomly, **Teacher Entropy** favors confident teacher predictions, and **Student–Teacher KL** prioritizes distributional disagreement. **High-Budget MC** uses additional continuations as a higher-fidelity counterfactual reference. Table 3 shows that random, entropy-, and disagreement-based alternatives trail counterfactual guidance. These results support evaluating teacher and student actions under the same continuation policy rather than relying on confidence or disagreement as utility proxies. The gap to High-Budget MC suggests further gains from more accurate estimation.

A stronger teacher is not locally better at every state. We test the key motivation behind CFOPD on a held-out set partitioned by problem identity, using 64 Monte Carlo continuations per branch under the same frozen student policy. As shown in Figure 3, despite the teacher's higher overall task accuracy, approximately 30%–50% of the evaluated states are *student-preferred*, where the student action attains a higher counterfactual value than the corresponding teacher intervention. Moreover, the learned gate suppresses approximately 79% of these student-preferred interventions while retaining more than 75% of teacher-preferred ones. These results show that global teacher superiority does not imply local action dominance, and that the counterfactual signal appropriately reduces teacher transfer where the student's own action is estimated to be more beneficial.

Does CF-PRM accurately predict intervention utility? We evaluate CF-PRM on held-out intervention pairs by comparing its predicted teacher advantages with high-budget references estimated using 64 Monte Carlo continuations per branch under the same frozen student policy. As shown in Figure 4, the predictions correlate well with the reference estimates, achieving a Pearson correlation of approximately 0.59 and a sign accuracy of 72.3%. Most pairs fall in the sign-consistent quadrants, indicating that CF-PRM captures both the relative magnitude and ordering of intervention utility. This distinction is critical for CFOPD, since the sign of the predicted advantage determines the direction of local counterfactual guidance and how strongly teacher information is transferred.

Online refresh tracks an evolving student policy. Because counterfactual values depend on the continuation ability of the current student, they may become stale as the policy evolves. We compare the default periodically refreshed CF-PRM with a **Frozen CF-PRM** trained only on initial counterfactual trajectories, evaluating both against high-budget references constructed from the current student policy using 64 Monte Carlo continuations per branch. As shown in Figure 5, the sign accuracy of Frozen CF-PRM steadily declines during training, whereas periodic refresh maintains substantially higher agreement with the reference action orderings. This confirms that counterfactual values are policy-dependent and motivates updating the value model alongside the evolving student.

Table 3: Comparison of adaptive supervision signals on MATH-500. High-Budget MC provides a higher-budget counterfactual reference.

Selection Signal	Avg@8	Pass@8
Uniform OPD	49.25	72.4
Random Matched	42.35	63.8
Teacher Entropy	50.43	66.2
Student–Teacher KL	48.38	71.8
CFOPD	**55.97**	**81.4**
High-Budget MC	57.93	83.6

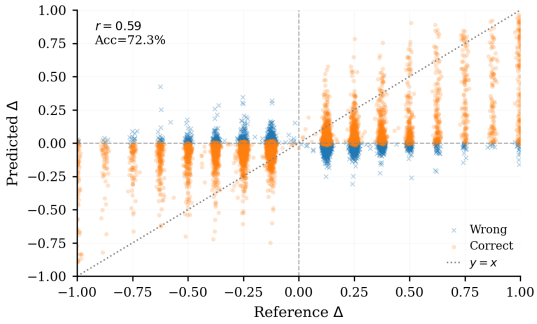

Figure 4: CF-PRM predictions versus high-budget counterfactual references on held-out pairs.

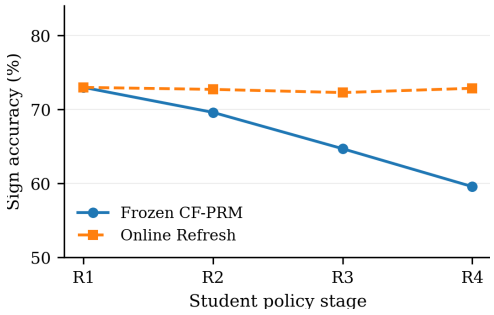

Figure 5: Performance of frozen versus periodically refreshed CF-PRM across training stages.

Table 4: Effect of the Monte Carlo rollout budget on counterfactual action ordering and downstream performance.

Rollouts / Branch	Sign Acc.	MATH-500 Avg@8
1	53.10	39.90
2	58.90	41.70
4	67.02	42.56
8	72.34	44.28
16	75.20	45.48

Computation–performance trade-off. The main computational overhead of CFOPD comes from counterfactual rollouts and value-model training. Table 4 shows the effect of varying the number of continuations per intervention branch. Increasing the rollout budget improves both counterfactual ordering accuracy and downstream reasoning performance, with diminishing gains at larger budgets. The improvement trends are consistent across different training settings, indicating that additional counterfactual samples provide more reliable intervention value estimation but introduce increasing computational cost. Since most of the improvement is achieved with eight continuations per branch, we use this setting by default to balance estimation quality and training cost.

5 Conclusion

We introduced CFOPD to evaluate teacher interventions relative to the current student's continuation policy. A shared counterfactual value model compares teacher and student actions at the same state to guide student optimization and provides process feedback along student trajectories. Across two teacher–student settings, CFOPD consistently improves reasoning performance over standard OPD. The analyses show that many local states favor the student action despite the teacher's higher overall accuracy, counterfactual utility is more informative than the tested heuristic signals, and value estimates benefit from periodic refresh as the student evolves. CFOPD is also more robust to corrupted teacher guidance.

These findings support adapting student optimization to policy-relative intervention utility rather than assuming uniformly beneficial teacher supervision. More broadly, effective on-policy distillation requires evaluating both the states where supervision is collected and the downstream utility of the guidance provided. CFOPD's additional computation is confined to training; the final student architecture and inference cost are unchanged.

AI USE STATEMENT

Generative AI tools were used during the preparation of this manuscript for assistance with language editing, proofreading, and improving the clarity of technical descriptions. They were also used to assist with minor code-related tasks, such as debugging and formatting.

Generative AI tools were not used to generate experimental results, design the research methodology, conduct scientific analysis, or make research claims. All AI-assisted outputs were reviewed and verified by the authors.

The authors take responsibility for the final content of this work, including the text, claims, code, experimental results, and other artifacts.

ETHICS STATEMENT

This work does not involve human subjects, private user data, or human-generated annotations. The datasets used in our experiments are publicly available benchmark datasets for mathematical reasoning. We follow the corresponding dataset licenses and use the resources solely for research purposes.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

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

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2402.14008.

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 *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Associ-

ation for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.07079.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.14858.

Enhan Li, Junhao He, and Hongyang Du. Crop: Task relevance via counterfactuals for selective on-policy distillation, 2026. URL https://arxiv.org/abs/2608.13387.

Wendi Li and Yixuan Li. Process reward model with q-value rankings. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/26494b66ae1114d314673e25b4967288-Abstract-Conference.html.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning, 2026. URL https://arxiv.org/abs/2605.21606.

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, 2024. URL https://arxiv.org/abs/2406.06592.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Khouloud Saadi and Di Wang. Validity-calibrated reasoning distillation, 2026. URL https://arxiv.org/abs/2605.04078.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/98711dea460bdefe0e651ca23ec98ba2-Abstract-Conference.html.

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. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation, 2026. URL https://arxiv.org/abs/2605.26844.

Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7309–7319. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.findings-emnlp.429. URL https://aclanthology.org/2024.findings-emnlp.429/.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation, 2026. URL https://arxiv.org/abs/2604.14084.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2066. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/76ec4dc30e9faaf0e4b6093eaa377218-Abstract-Conference.html.

Zhiwei Zhang, Zechen Sun, Fei Zhao, Kang Peng, Bin Liang, Huayu Deng, Yao Hu, Kam-Fai Wong, and Mu Chuan. Verify before you distill: Prompt-level teacher gating for on-policy distillation, 2026. URL https://arxiv.org/abs/2609.02998.

A Dataset Details

We evaluate CFOPD on mathematical reasoning tasks using two training settings with different teacher–student configurations. This section provides details of the training datasets and evaluation benchmarks used in our experiments.

A.1 Training Datasets

MATH. The MATH dataset (Hendrycks et al., 2021) is used in the MATH-trained setting. It consists of challenging mathematical problems collected from competition sources, covering various areas including algebra, geometry, number theory, and probability. Each problem contains a final answer and a detailed solution process, making it a widely used benchmark for evaluating mathematical reasoning models.

GSM8K. The GSM8K dataset (Cobbe et al., 2021) is used in the GSM8K-trained setting. It contains grade-school mathematical word problems that require multi-step reasoning to derive the final answer. Compared with MATH, GSM8K focuses more on elementary mathematical reasoning and numerical problem solving.

A.2 Evaluation Benchmarks

For the MATH-trained setting, we evaluate models on MATH-500, AMC23, Minerva Math (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), AIME24, and AIME25. These benchmarks cover a broad range of mathematical reasoning difficulties, including general problem solving, competition mathematics, and olympiad-level reasoning.

MATH-500. MATH-500 is a commonly used subset of the MATH benchmark for evaluating mathematical reasoning ability.

AMC23. AMC23 consists of problems from the American Mathematics Competition and evaluates mathematical problem-solving ability on competition-style questions.

Minerva Math. Minerva Math (Lewkowycz et al., 2022) contains challenging mathematical problems designed to test advanced reasoning capabilities.

OlympiadBench. OlympiadBench (He et al., 2024) provides olympiad-level mathematical problems for evaluating long-horizon reasoning.

AIME24 and AIME25. AIME24 and AIME25 contain highly challenging mathematics competition problems and are used to measure generalization to difficult reasoning tasks.

For the GSM8K-trained setting, we evaluate models on the same six benchmarks used in the MATH-trained setting.

A.3 Evaluation Protocol

All methods are evaluated using the same decoding and verification settings. For each evaluation problem, we independently sample eight responses. We report Avg@8, which measures the average correctness over sampled responses, and Pass@8, which measures whether at least one sampled response obtains the correct answer.

B Additional Experimental Details

Training configuration. All methods use the same student initialization and optimization budget within each teacher–student setting. Student trajectories are sampled with temperature 0.7, with maximum response lengths of $2{,}048$ tokens for GSM8K and $4{,}096$ tokens for MATH. We use a training batch size and mini-batch size of 16, and optimize the student with a fixed learning rate of

Table 5: Complementarity of action-level guidance and gated distillation on MATH-500.

Method	Avg@8	Pass@8
Action Guidance only	53.93	77.60
Gated Distillation only	53.05	75.80
CFOPD	55.97	81.40

Table 6: Robustness to corrupted teacher guidance. Teacher reasoning steps are replaced with randomly sampled token sequences during counterfactual supervision construction and CF-PRM training. Results are reported under strict and relaxed evaluation protocols.

Method	**GSM8K**				**MATH-500**			
	Strict		**Relaxed**		**Strict**		**Relaxed**	
	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8	Avg@8	Pass@8
GKD-OPD	4.38	16.60	7.49	22.59	0.0	0.0	1.25	8.60
CFOPD	**9.24**	**26.84**	**17.83**	**35.71**	**3.4**	**9.2**	**5.48**	**13.4**

2×10^{-6}. Unless otherwise specified, training is performed for three epochs. All methods use the same prompt formatting, answer extraction, and verification procedure within each setting.

Counterfactual supervision. For CFOPD, counterfactual supervision is constructed at up to four selected reasoning steps per trajectory. At each selected state, the teacher provides a one-step intervention, after which both branches are continued by the same frozen student policy. Each branch uses eight Monte Carlo continuations by default. The counterfactual value model is updated for one epoch at each refresh with a learning rate of 1×10^{-5}. We set $\lambda_{\text{eff}} = 5.0$, $\lambda_{\text{rank}} = 3.0$, $\alpha = 1.0$, $\lambda_{\text{pref}} = 1.0$, and $\beta = 1.0$, shared across settings. Counterfactual targets and the value model are periodically refreshed using trajectories from the updated student.

Implementation and hardware. Our implementation uses the VERL framework for distributed training infrastructure and vLLM for efficient response generation. Training and inference use bfloat16 precision. Experiments are conducted on NVIDIA RTX 3090 GPUs and NVIDIA RTX 5880 GPUs , with counterfactual generation and evaluation parallelized across the available devices. The same hardware configuration is used for the reported training-time comparisons.

C Additional Experimental Results

Are the two uses of counterfactual guidance complementary? We separate the two uses of the counterfactual action contrast. *Action Guidance only* updates the evaluated teacher–student action pair, whereas *Gated Distillation only* uses the contrast only to modulate teacher distribution transfer. Full CFOPD combines both. As shown in Table 5, each use improves performance individually, while their combination performs best, supporting their complementary roles.

Robustness to corrupted teacher guidance. We evaluate CFOPD under corrupted teacher guidance by replacing teacher reasoning steps with randomly sampled token sequences when constructing counterfactual supervision. The resulting corrupted teacher-derived signals are used consistently throughout training, while all other settings remain unchanged. Table 6 reports performance under strict and relaxed evaluation protocols.

Both GKD-OPD and CFOPD degrade under corrupted guidance, showing that unreliable teacher signals can adversely affect on-policy distillation. Nevertheless, CFOPD consistently outperforms GKD-OPD on both GSM8K and MATH-500 under strict and relaxed evaluation protocols. These results suggest that policy-relative counterfactual guidance provides additional robustness when teacher supervision is imperfect.

Table 7: Wall-clock training time under the same hardware configuration.

Method	student Training	Auxiliary Supervision	Total
PG-OPD	3.1 h	–	3.1 h
GKD-OPD	3.3 h	–	3.3 h
CFOPD	3.3 h	3.2 h	6.5 h

Auxiliary supervision includes counterfactual label construction and value-model training.

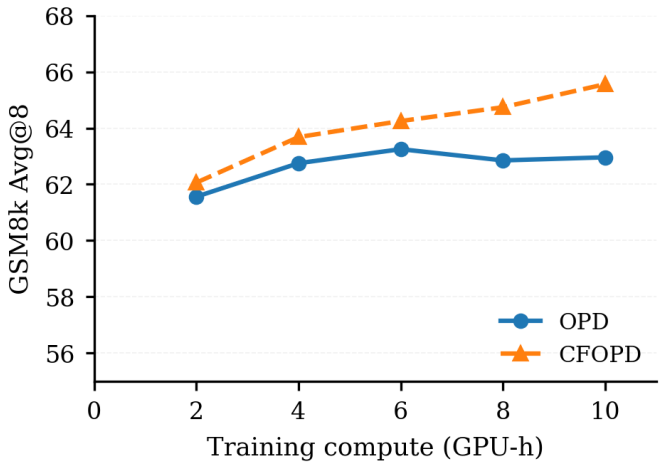

Figure 6: **Computation–performance trade-off of CFOPD.** GSM8K Avg@8 versus total training compute for CFOPD and standard OPD under increasing optimization budgets.

D Efficiency Analysis

Training efficiency. Table 7 reports wall-clock training time under the same hardware configuration. CFOPD requires 6.5 hours in total, compared with 3.1 and 3.3 hours for PG-OPD and GKD-OPD, respectively. Its student optimization cost remains close to GKD-OPD, with the dominant additional overhead coming from counterfactual supervision construction and auxiliary value-model training. This cost is incurred only during training and does not affect the final student architecture or inference cost.

Computation–performance trade-off. Figure 6 examines whether the gains of CFOPD can be explained simply by increased training compute. We compare CFOPD with standard OPD under progressively expanded optimization budgets. Although additional compute improves the performance of OPD, CFOPD achieves a more favorable computation–performance trade-off and remains stronger at comparable training costs. This suggests that the improvement of CFOPD arises primarily from more effective counterfactual guidance rather than from increased training compute alone.

PRM-free CFOPD. Although CFOPD provides effective counterfactual supervision, its additional cost mainly arises from Monte Carlo continuation labeling and repeated training of the counterfactual value model. To reduce this overhead, we develop a PRM-free variant that replaces the explicit CF-PRM with an implicit potential computed by a frozen snapshot of the student. Given a reasoning state s and the canonical gold answer $y^* = (y_1^*, \ldots, y_m^*)$, we evaluate the length-normalized likelihood of the answer under teacher forcing:

$$\Phi(s) = \exp\left(\frac{1}{m}\sum_{j=1}^{m}\log \bar{\pi}\left(y_j^* \mid s, y_{<j}^*\right)\right), \tag{15}$$

where $\bar{\pi}$ is a frozen student snapshot. For a candidate action a, we use $\Phi(s \oplus a)$ as an implicit action value in place of $q_\phi(s, a)$. The same temporal and action contrasts as in CFOPD are then constructed

Table 8: Performance and training efficiency of the PRM-free CFOPD variant on MATH-500.

Method	Avg@8	Pass@8	Training Time
GKD-OPD	50.65	75.20	3.3 h
CFOPD	55.97	81.40	6.5 h
CFOPD-Implicit	53.95	79.20	3.4 h

from these implicit values. This removes explicit counterfactual rollout labeling and value-model optimization, substantially reducing auxiliary supervision cost while retaining the policy-relative comparison between student and teacher actions.

As shown in Table 8, under the MATH-trained setting, CFOPD-Implicit substantially reduces the additional training overhead introduced by explicit counterfactual labeling and CF-PRM optimization, while preserving a large portion of the performance improvement over standard OPD. Although full CFOPD achieves stronger overall performance, the PRM-free variant provides a more efficient accuracy–cost trade-off, suggesting that policy-relative counterfactual guidance can be retained without explicitly training a learned value model.