

Consistency-Calibrated Advantage Estimation for Label-Free Reinforcement Learning

Meng Yang^{1*}, Yongle Zhao^{2*}, Zhiyang Zhang¹, Xiaorui Jiang¹,
Peirou Liang¹, Furui Chen¹, Yin Xie², Ziyong Feng^{2†}

¹University of Science and Technology of China, Hefei, China

²DeepGlint, Beijing, China

Correspondence: ziyongfeng@deepglint.com

Abstract

Reinforcement learning from unlabeled prompts offers a promising path to improving LLM reasoning without curated supervision or verifiable rewards. However, existing label-free methods typically rely on either intrinsic self-certainty or extrinsic answer agreement, both of which can become unreliable in hard reasoning regimes: confident trajectories can be wrong, while majority answers can drift toward stable yet incorrect attractors. We propose **CCA**E, a verifier-free and parameter-free advantage estimator that formulates label-free GRPO as a *multi-signal advantage estimation* problem. **CCA**E decomposes the advantage into a base self-certainty term, an intrinsic evidence factor, and an extrinsic agreement factor, with their interaction adaptively calibrated by a group-level proxy for agreement reliability, without introducing any reward model, verifier, or trainable parameters. On MATH-Train, **CCA**E improves over the strongest unlabeled baseline by 5.0 and 4.3 points on Qwen3-4B/8B-Base, respectively, matching supervised GRPO. On the harder DAPO-14k-MATH, **CCA**E remains stable while prior unlabeled methods collapse, outperforming their best pre-collapse checkpoints by up to 11.1 points. Results on Natural Reasoning further suggest that **CCA**E remains effective beyond math-only training. Code is available at: <https://github.com/tifing/CCA>.

1 Introduction

Recent advances in large language models (LLMs) have yielded strong performance on challenging reasoning tasks (Shao et al., 2024; Snell et al., 2025). However, pushing beyond supervised learning still relies heavily on post-training with reinforcement learning (RL) (Ziegler et al., 2019; Ouyang et al., 2022; Rafailov et al., 2023). In

particular, RL with verifiable rewards (RLVR) can substantially improve mathematical and coding reasoning when an automatic verifier is available (Guo et al., 2025; Shao et al., 2025). However, RLVR remains inherently constrained by the availability and reliability of verifiers: in open-ended or newly emerging domains, verifiers are often missing, expensive, or brittle.

These limitations motivate unlabeled self-improvement, which leverages abundant questions and prompts without direct supervision (Zhao et al., 2025). The central challenge is not the absence of pseudo-signals, but their reliability. Existing approaches primarily rely on either intrinsic confidence criteria (e.g., entropy regularization or *self-certainty*) (Grandvalet and Bengio, 2004; Wang et al., 2020; Agarwal et al., 2025; Prabhudesai et al., 2025; Kang et al., 2025) or agreement-based signals (e.g., self-consistency and majority voting) that yield pseudo-labels (Lee et al., 2013; Wang et al., 2022; Chen et al., 2023; Zuo et al., 2025).

Despite their appeal, both signal families exhibit complementary failure modes in challenging reasoning regimes. Internal self-certainty can reinforce confident but incorrect trajectories and drive entropy collapse, while external consensus can decouple from correctness on hard tasks and drift toward stable yet incorrect attractors. Consequently, directly relying on either signal, or naively combining the two without accounting for their reliability, can trigger self-reinforcing *dynamics collapse*, characterized by sustained accuracy degradation, reduced exploration, and increasingly brittle model behavior over optimization (Shumailov et al., 2024; Gao et al., 2023).

To address this, we introduce **Consistency-Calibrated Advantage Estimation (CCA**E), a verifier-free and parameter-free advantage estimator for label-free reinforcement learning. Rather than directly optimizing either pseudo-signal,

*Equal contribution.

†Corresponding author.

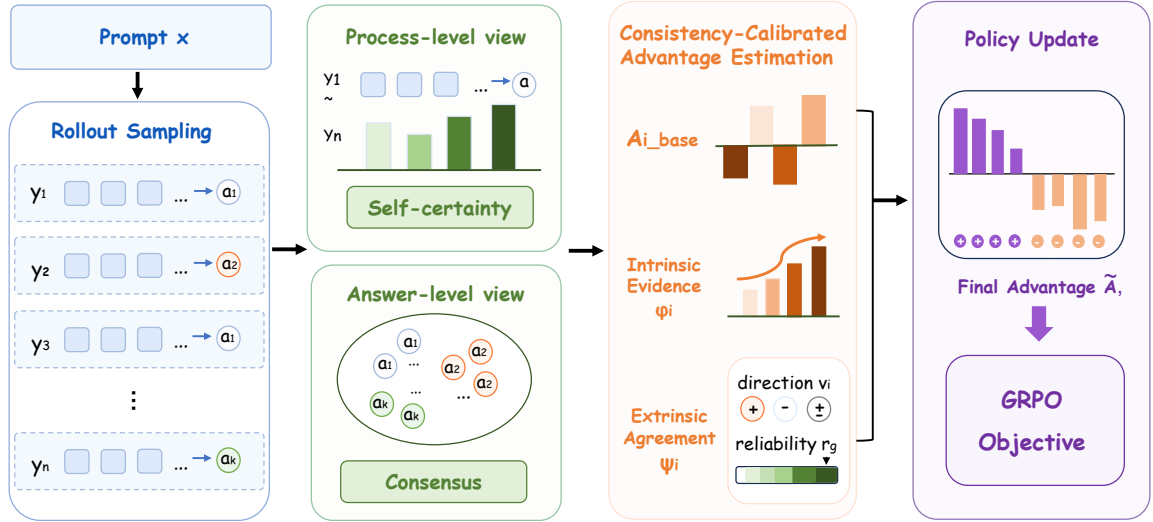


Figure 1: Overview of **CCA**E. For each prompt, the policy samples a group of reasoning rollouts. **CCA**E derives label-free training signals from two views: process-level self-certainty over the generated reasoning trajectory and answer-level agreement among final responses. A consistency-calibrated advantage estimator then combines three factors: the base self-certainty advantage, the intrinsic evidence factor, and the extrinsic agreement factor. The calibrated final advantage is used to optimize the policy with GRPO.

CCAE formulates label-free learning as a multi-signal advantage estimation problem by leveraging two complementary views: process-level self-certainty over the reasoning trajectory and answer-level agreement among final responses. **CCA**E decomposes the final advantage into a base self-certainty advantage, an intrinsic evidence factor, and an extrinsic agreement factor, allowing the two signals to play complementary roles in policy optimization. Group-level consensus adaptively calibrates how these signals influence the update, without introducing reward models, learned verifiers, hidden-state features, or additional trainable parameters.

Our contributions are threefold:

- We recast label-free reinforcement learning as a multi-signal advantage estimation problem using process-level self-certainty and answer-level agreement, two complementary but individually unreliable pseudo-signals.
- We propose **CCA**E, a verifier-free and parameter-free advantage estimator that decomposes the advantage into a base self-certainty term, an intrinsic evidence factor, and an extrinsic agreement factor, with group-level consensus adaptively calibrating their influence. **CCA**E requires no reward model, learned verifier, hidden-state features, or additional trainable parameters.

- **CCA**E consistently outperforms strong unlabeled baselines across mathematical and non-mathematical reasoning settings and a non-Qwen backbone, while remaining stable on the harder DAPO-14k-MATH setting, where prior methods collapse, with gains of up to 11.1 points over their best pre-collapse checkpoints.

2 Related Work

RLHF and preference optimization have been widely used to improve LLM behaviors, but they typically depend on human annotations or learned reward models (Ziegler et al., 2019; Ouyang et al., 2022; Rafailov et al., 2023). In reasoning-centric domains, RL with verifiable rewards (RLVR) further replaces learned rewards with automatic checks (e.g., exact match or execution), and commonly adopts PPO-style updates and group-based objectives (Schulman et al., 2017; Ahmadian et al., 2024; Shao et al., 2024; Guo et al., 2025; Hu et al., 2025). Our focus differs: we study label-free self-learning settings where such external supervision is unavailable or too costly to obtain.

Entropy minimization and confidence-driven objectives have a long history in semi-supervised learning and test-time adaptation (Grandvalet and Bengio, 2004; Wang et al., 2020), and have recently been revisited for unlabeled LLM reasoning (Agarwal et al., 2025; Prabhudesai et al., 2025).

A representative family of signals is *self-certainty* (e.g., $\text{KL}(U \| p_\theta)$), which provides a scalable, fully internal training signal without external rewards (Kang et al., 2025). However, these signals are not inherently correctness-aware: mode-seeking confidence optimization can accelerate entropy collapse and amplify high-confidence errors, especially in hard reasoning regimes.

CCAE targets label-free self-learning where both internal confidence and external consensus can be unreliable. Rather than directly optimizing either pseudo-signal, **CCAE** formulates training as a multi-signal advantage estimation problem. It uses process-level self-certainty as intrinsic evidence and answer-level agreement for outcome-level calibration, yielding a verifier-free and parameter-free estimator that can be directly inserted into GRPO.

3 Consistency-Calibrated Advantage Estimation

CCAE is a verifier-free and parameter-free advantage estimator for label-free reinforcement learning. It does not introduce an auxiliary reward model, a learned verifier, hidden-state features, or extra trainable parameters. Instead, it changes only how the final GRPO advantage is estimated.

The central problem is that, in label-free RL, the true advantage of a rollout is unobserved. Existing methods usually rely on a single pseudo-signal. Self-certainty based methods use the model’s own confidence as a dense intrinsic signal, but confidence does not specify whether a rollout should be reinforced. Voting based methods use answer agreement as an outcome-level signal, but the majority answer may be wrong in hard regimes. Thus, self-certainty provides useful magnitude but lacks reliable direction, while answer agreement provides useful direction but can be unreliable alone.

We formulate label-free RL as a multi-signal advantage estimation problem. Given a group of rollouts, **CCAE** estimates each rollout’s advantage by combining three factors:

$$\tilde{A}_i = A_i^{\text{base}} \cdot \phi_i \cdot \psi_i. \quad (1)$$

Here A_i^{base} is the base advantage derived from self-certainty, ϕ_i measures whether the intrinsic signal is sufficiently salient, and ψ_i calibrates the update direction using answer agreement.

3.1 Label-Free GRPO Setup

For each prompt x , the policy π_θ samples a group of G rollouts:

$$\{y_i\}_{i=1}^G \sim \pi_\theta(\cdot | x). \quad (2)$$

Each rollout y_i contains a reasoning trajectory and a final answer. Ground-truth rewards are not used during training. The goal is to construct a label-free advantage estimate for each rollout and use it in the standard GRPO update.

Let $\tilde{A}_i$ denote the final advantage used by the policy loss. **CCAE** modifies only the estimator that produces $\tilde{A}_i$. The rollout procedure, clipped GRPO objective, KL regularization, and optimizer remain unchanged.

3.2 Self-Certainty as the Base Intrinsic Signal

For rollout $y_i = (y_{i,1}, \dots, y_{i,T_i})$, let $p_\theta(\cdot | x, y_{i,<t})$ denote the next-token distribution at step t . We compute the self-certainty score as

$$s_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \text{KL}(U \| p_\theta(\cdot | x, y_{i,<t})), \quad (3)$$

where U is the uniform distribution over the vocabulary. A larger s_i indicates a sharper predictive distribution along the generated trajectory.

Within each rollout group, we standardize self-certainty:

$$z_i = \frac{s_i - \mu_s}{\sigma_s + \epsilon}, \quad (4)$$

where μ_s and σ_s denote the mean and standard deviation of self-certainty scores within the rollout group.

The base self-certainty advantage is

$$A_i^{\text{base}} = z_i. \quad (5)$$

This is the standard GRPO-style normalized advantage obtained from the intrinsic signal. It captures relative confidence within the prompt group. However, it is not directionally reliable by itself, since confident trajectories can still be wrong.

3.3 Answer Agreement as an Extrinsic View

We extract the final answer $\hat{a}_i$ from each rollout and cluster rollouts by answer equivalence. Let p_{a_i} be the fraction of rollouts that produce the same answer as rollout i :

$$p_{a_i} = \frac{|\{j : \hat{a}_j = \hat{a}_i\}|}{G}. \quad (6)$$

Let $p_{\max} = \max_i p_{a_i}$ be the largest answer-cluster proportion, and let K_g be the number of distinct answer clusters in the group. We define the uniform-support baseline as

$$p_0 = \frac{1}{K_g}. \quad (7)$$

The group-level reliability of the agreement signal is

$$r_g = \frac{p_{\max} - p_0}{1 - p_0 + \epsilon}. \quad (8)$$

If all rollouts fall into a single answer cluster, we set $r_g = 1$. When answers are uniformly spread across clusters, r_g is close to 0. When one answer dominates the group, r_g is close to 1.

For each rollout, we define a vote-direction score:

$$v_i = \text{clip} \left(\frac{p_{a_i} - p_0}{p_{\max} - p_0 + \epsilon}, -1, 1 \right). \quad (9)$$

Thus, rollouts supported by high-frequency answers receive larger v_i , while weakly supported answers receive smaller or negative values. The role of v_i is to provide outcome-level directional evidence.

3.4 Consistency-Weighted Advantage Estimator

CCAE estimates the final advantage by multiplying the base intrinsic advantage with two consistency factors. The first factor measures the strength of intrinsic evidence:

$$\phi_i = |z_i| \cdot \mathbf{1}[z_i > 1 - r_g]. \quad (10)$$

The magnitude $|z_i|$ reflects how strongly rollout i deviates from the group baseline under the self-certainty view. The indicator represents evidence sufficiency. When answer agreement is reliable, r_g is large and the condition becomes easier to satisfy. When agreement is weak, r_g is small and the estimator requires stronger intrinsic evidence.

The second factor incorporates extrinsic directional calibration:

$$\psi_i = (1 - r_g) + r_g v_i. \quad (11)$$

When r_g is small, answer agreement is unreliable and ψ_i is close to 1. In this case, the estimator falls back toward self-certainty based learning. When r_g is large, agreement becomes more informative and ψ_i is mainly determined by the vote-direction score v_i .

The final advantage estimate is

$$\tilde{A}_i = A_i^{\text{base}} \cdot \phi_i \cdot \psi_i. \quad (12)$$

This equation is the core of **CCA**E. The base term provides the initial self-certainty preference. The intrinsic factor controls whether the self-certainty signal is salient enough to contribute. The extrinsic factor calibrates the update direction using answer agreement. Therefore, high self-certainty alone is not sufficient to produce a strong positive update. The final update also depends on whether the rollout is consistent with outcome-level evidence.

3.5 Relation to Single-Signal Estimators

The estimator in Eq. (12) generalizes single-signal label-free advantage estimators.

If answer agreement is uninformative, then $r_g \approx 0$ and $\psi_i \approx 1$. The estimator is then mainly determined by the self-certainty advantage and the intrinsic evidence factor. This corresponds to intrinsic-signal learning.

If answer agreement is highly concentrated, then $r_g \approx 1$ and $\psi_i \approx v_i$. The estimator then uses agreement to calibrate the direction of the update, while self-certainty still controls the strength of the signal through A_i^{base} and ϕ_i .

Thus, **CCA**E does not simply add voting to self-certainty. It decomposes label-free advantage estimation into base preference, intrinsic evidence, and extrinsic directional calibration. This makes the final advantage depend on multi-signal consistency rather than on any single pseudo-signal alone.

3.6 Training Objective

We replace the standard GRPO advantage with $\tilde{A}_i$. The clipped GRPO objective becomes

$$\begin{aligned} \mathcal{L}_{\text{CCA}}(x; \theta) = & \frac{1}{G} \sum_{i=1}^G \min \left(\rho_i(\theta) \tilde{A}_i, \right. \\ & \left. \text{clip}(\rho_i(\theta), 1 - \epsilon, 1 + \epsilon) \tilde{A}_i \right) \\ & - \beta \text{KL}(\pi_\theta \parallel \pi_{\text{ref}}). \end{aligned} \quad (13)$$

The policy optimization procedure is otherwise unchanged. **CCA**E only changes the advantage estimator and can be inserted into existing GRPO training pipelines.

Table 1: **Main results on MATH-Train with Qwen3-4B-Base and Qwen3-8B-Base.** Pass@1 (%) is reported on six reasoning benchmarks. All results are averaged over 16 generations. The best results among unlabeled methods are highlighted in **bold**. **CCAE** consistently achieves the highest average performance among unlabeled baselines and remains competitive with supervised GRPO trained with gold labels.

Model	MATH500	AIME25	AMC23	Olym	MMLU	GPQA-d	Avg. ↑
Qwen3-4B-Base	67.8	6.3	43.1	33.4	60.0	29.9	40.1
<i>w/ access to gold label</i>							
GRPO	81.5	17.5	56.5	44.6	74.2	37.5	52.0
<i>w/o access to gold label</i>							
INTUITOR	75.3	10.6	48.6	37.9	74.7	35.9	47.2
TTRL	79.3	12.9	50.9	41.0	59.8	33.6	46.3
EMPO	75.9	14.6	50.3	39.3	61.0	33.7	45.8
CCAE (Ours)	79.9	16.7	56.9	42.1	79.2	38.3	52.2
Qwen3-8B-Base	63.9	10.8	45.6	32.1	66.6	32.4	41.9
<i>w/ access to gold label</i>							
GRPO	84.4	18.1	61.3	46.8	73.0	42.3	54.3
<i>w/o access to gold label</i>							
INTUITOR	78.3	9.4	55.0	39.4	74.5	38.5	49.2
TTRL	81.9	15.4	59.7	44.1	62.9	37.9	50.3
EMPO	81.5	15.6	59.8	43.6	62.8	38.5	50.3
CCAE (Ours)	82.9	17.5	61.6	45.1	77.6	42.7	54.6

4 Experimental Setup

Training Data. We consider two mathematical reasoning corpora with different scales and difficulty profiles, together with an additional natural reasoning corpus for evaluating generalization beyond mathematics.

- **MATH-Train-7.5k** (Hendrycks et al., 2021). This split contains 7,500 problems and represents a moderate-difficulty training regime commonly used in prior work.
- **DAPO-14k-MATH.** This dataset is derived from DAPO-Math-17k (Yu et al., 2025) by deduplicating prompts and standardizing prompt and answer formats. We exclude approximately 3k non-English prompts and retain 14k English problems without further modification. Compared to MATH-Train-7.5k, DAPO-14k-MATH contains a larger proportion of harder problems and induces more unstable training dynamics under unsupervised self-learning.
- **Natural Reasoning.** We use Natural Reasoning as a non-mathematical reasoning corpus to study whether label-free training generalizes beyond math-only prompts. Compared with the mathematical datasets, its responses are more free-

form and less suitable for exact symbolic verification, making it a complementary setting for testing agreement-based and confidence-based pseudo-signals.

Base Models. We evaluate **CCAE** on three base models spanning different families and scales: **Qwen3-4B-Base** (trained on both MATH-Train-7.5k and DAPO-14k-MATH to study training difficulty under a fixed scale), **Qwen3-8B-Base** (trained on MATH-Train-7.5k to study the effect of model scale under the same training setting), and **OctoThinker Hybrid-8B-Base** (Wang et al., 2025) (a non-Qwen backbone, trained on MATH-Train-7.5k).

Training Setup. Within each base model, all methods share the same RL pipeline and optimization hyperparameters. **CCAE** adds no trainable parameters or tuned thresholds; all quantities are computed from the rollout group. Full hyperparameters are listed in Appendix A.1.

Benchmarks and Evaluation. We evaluate on six public benchmarks covering math reasoning (**MATH-500** (Hendrycks et al., 2021), **AIME25** (Zhang and Team, 2025), **AMC23**, **Olympiad-**

Bench math subset (He et al., 2024)) and STEM reasoning (MMLU_STEM, GPQA-Diamond). For non-mathematical reasoning, we additionally use GPQA and MMLU-Pro. For each problem, we sample 16 completions (temperature 0.6, top- p 0.95) and report average Pass@1, following the Qwen2.5-math evaluation codebase (Yang et al., 2024).

Baselines. We compare CCAE with three label-free baselines: INTUITOR (Zhao et al., 2025) uses self-certainty as an intrinsic reward; TTRL (Zuo et al., 2025) constructs pseudo-labels via majority voting over rollouts; and EMPO (Zhang et al., 2025) minimizes semantic entropy across sampled solutions. See Appendix A.3 for details.

5 Results and Analysis

5.1 Results on MATH-Train

Table 1 reports the main results on MATH-Train. Across both model scales, CCAE achieves the best average Pass@1 among all unlabeled methods (52.2% on Qwen3-4B-Base, 54.6% on Qwen3-8B-Base), improving over the strongest unlabeled baseline by +5.0 and +4.3 points respectively, with consistent gains across all six benchmarks and average improvements of +12.1 and +12.7 points over the base models.

Notably, CCAE achieves consistent gains across both mathematical (MATH-500, AIME25, AMC23, OlympiadBench) and broader reasoning benchmarks (MMLU_STEM, GPQA-Diamond). Despite using no gold labels, CCAE matches supervised GRPO in average performance and even exceeds it on MMLU_STEM and GPQA-Diamond, suggesting that calibrating self-certainty with answer-level agreement yields effective advantage estimates for label-free RL.

5.2 Results on DAPO-14k-MATH

DAPO-14k-MATH provides a more challenging label-free reinforcement learning setting than MATH-Train. In this regime, pseudo-signals are noisier and prior unlabeled methods become more prone to unstable optimization. In our experiments, INTUITOR, TTRL, and EMPO all suffer from training collapse. Therefore, Table 2 reports their best pre-collapse checkpoints selected using training accuracy with gold labels. This gives these baselines an optimistic comparison, since such checkpoint selection is not available in a truly

label-free setting. In contrast, CCAE reports the final checkpoint without any gold-label checkpoint selection.

Main results. As shown in Table 2, CCAE achieves the best average performance among unlabeled methods, reaching an average Pass@1 of 52.1%. Even compared with the optimistic pre-collapse checkpoints of prior baselines, CCAE outperforms INTUITOR, TTRL, and EMPO by +3.1, +6.6, and +11.1 points on average, respectively. Compared with the base model, CCAE improves the average Pass@1 from 40.1% to 52.1%, yielding a gain of +12.0 points. CCAE also achieves the best unlabeled result on all six benchmarks, showing consistent improvements across both mathematical reasoning and broader reasoning tasks.

Training dynamics. Figure 2 analyzes the training dynamics on DAPO-14k-MATH. Prior unlabeled methods exhibit clear instability: confidence-based learning aggressively collapses policy entropy, while agreement-based learning accumulates erroneous updates as the majority answer becomes unreliable. In contrast, CCAE maintains a substantially lower erroneous update ratio and higher voting accuracy, with policy entropy and response length remaining bounded. These dynamics explain why CCAE can report the final checkpoint on DAPO-14k-MATH without gold-label-based checkpoint selection, while prior baselines require optimistic pre-collapse selection.

5.3 Performance on Natural Reasoning Tasks

To test generalization beyond math, we train CCAE and all baselines on the Natural Reasoning corpus and evaluate on natural reasoning benchmarks (GPQA and MMLU-Pro). Unlike mathematical training data with deterministic answers, Natural Reasoning contains more open-ended responses and is less amenable to exact symbolic verification. We therefore implement **Pseudo-TTRL**, a TTRL variant where rollouts are clustered by answer equivalence and the largest cluster is used as the pseudo label.

As shown in Table 3, CCAE achieves the strongest overall performance among unlabeled methods after training on Natural Reasoning. On GPQA, CCAE reaches 33.6%, outperforming INTUITOR, Pseudo-TTRL, and EMPO by +1.7,

Table 2: **Main results on DAPO-14k-MATH with Qwen3-4B-Base.** Pass@1 (%) is reported on six reasoning benchmarks. All results are averaged over 16 generations. The best results among unlabeled methods are highlighted in **bold**.

Model	MATH500	AIME25	AMC23	Olym	MMLU	GPQA-d	Avg. ↑
Qwen3-4B-Base	67.8	6.3	43.1	33.4	60.0	29.9	40.1
<i>w/ access to gold label</i>							
GRPO	79.6	21.3	59.2	45.0	81.1	38.5	54.1
<i>w/o access to gold label</i>							
INTUITOR	78.3	11.7	50.0	38.7	77.0	38.5	49.0
TTRL	77.2	6.7	51.3	38.6	63.7	35.4	45.5
EMPO	67.5	9.2	46.3	33.9	58.5	30.4	41.0
CCAE (Ours)	79.5	14.7	56.3	42.8	80.0	39.4	52.1

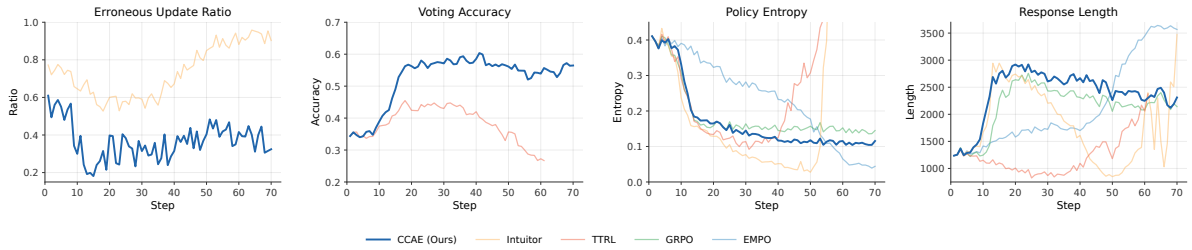


Figure 2: **Training dynamics on DAPO-14k-MATH.** We track erroneous update ratio, voting accuracy, policy entropy, and response length during label-free RL training. Compared with prior unlabeled baselines, **CCAE** maintains lower erroneous update ratio, higher voting accuracy, and more stable entropy and length dynamics, suggesting that reliability-calibrated advantage estimation mitigates noisy consensus and stabilizes policy updates.

Table 3: **Results after training on Natural Reasoning with Qwen3-4B-Base.** Pass@1 (%) averaged over 16 generations. For MMLU-Pro, the Avg. is the official overall accuracy across all categories. Best unlabeled results in **bold**.

Model	GPQA	MMLU-Pro				
		STEM	Hum.	Social	Other	Avg.
Qwen3-4B-Base	28.0	48.9	27.7	52.0	45.6	45.1
<i>w/ access to gold label</i>						
GRPO	34.8	60.2	33.6	65.2	53.9	55.8
<i>w/o access to gold label</i>						
INTUITOR	31.9	60.5	32.0	62.2	53.0	48.5
Pseudo-TTRL	32.5	52.7	33.2	61.6	50.0	50.1
EMPO	32.6	54.3	33.0	62.0	50.8	51.1
CCAE (Ours)	33.6	62.9	32.4	62.9	55.8	56.4

+1.1, and +1.0 points, respectively. On MMLU-Pro, **CCAE** obtains the best average score of 56.4%, improving over the strongest unlabeled baseline EMPO by +5.3 points.

We note that Pseudo-TTRL and EMPO suffer from training collapse in this setting. Therefore, the reported results for these two baselines correspond to their best pre-collapse checkpoints se-

lected with gold-label feedback. This provides an optimistic comparison and is not available in a truly label-free setting. In contrast, **CCAE** reports the final checkpoint without gold-label checkpoint selection.

The gains are consistent across most MMLU-Pro categories, with **CCAE** leading on STEM, Social Sciences, and Other (Pseudo-TTRL is marginally better on Humanities). These results demonstrate that **CCAE** extends beyond math-only training and remains effective on more open-ended natural reasoning data.

5.4 Additional Results on Broader Model Family beyond Qwen3

To further validate the generality of **CCAE** beyond the Qwen3 model family, we conduct additional experiments on OctoThinker Hybrid-8B-Base (Wang et al., 2025), a hybrid reasoning model built upon the Llama-3 model family through mid-training. The model is trained on MATH-Train under the same label-free setting.

As shown in Table 4, **CCAE** improves the average Pass@1 from 21.6% to 36.3%, yield-

Table 4: **Additional results on MATH-Train with OctoThinker Hybrid-8B-Base.** Pass@1 (%) is reported on six reasoning benchmarks. All results are averaged over 16 generations. The best results among unlabeled methods are highlighted in **bold**.

Model	MATH500	AIME25	AMC23	Olym	MMLU	GPQA-d	Avg. $\uparrow$
OctoThinker Hybrid-8B-Base	38.4	0.6	17.7	10.7	39.7	22.5	21.6
<i>w/ access to gold label</i>							
GRPO	68.0	7.7	44.2	31.4	53.1	24.2	38.1
<i>w/o access to gold label</i>							
INTUITOR	59.3	6.3	35.0	23.9	42.2	22.8	31.6
TTRL	61.6	6.5	34.7	24.1	53.0	23.2	33.9
CCAE (Ours)	63.9	7.1	37.6	28.5	56.2	24.6	36.3

Table 5: **Ablation study on MATH-Train and DAPO-14k-MATH.** Pass@1 (%) is reported on six reasoning benchmarks. All results are averaged over 16 generations. We ablate the intrinsic evidence factor ϕ_i , the extrinsic agreement rectification factor ψ_i , and the adaptive reliability factor r_g . The best results are highlighted in **bold**.

Training Data	Method	MATH500	AIME25	AMC23	Olym	MMLU	GPQA-d	Avg. $\uparrow$
MATH	CCAE (Full)	79.9	16.7	56.9	42.1	79.2	38.3	52.2
	- w/o Intrinsic Evidence ϕ_i	67.7	17.1	49.2	34.9	59.9	31.7	43.4
	- w/o Agreement Rectification ψ_i	78.5	15.8	54.9	40.1	79.2	37.1	50.9
	- w/o Adaptive Reliability r_g	78.9	16.5	53.5	42.3	78.4	37.6	51.2
DAPO	CCAE (Full)	79.5	14.7	56.3	42.8	80.0	39.4	52.1
	- w/o Intrinsic Evidence ϕ_i	59.7	16.5	46.9	31.8	53.3	31.8	40.0
	- w/o Agreement Rectification ψ_i	76.3	11.7	47.5	39.1	76.9	34.9	47.7
	- w/o Adaptive Reliability r_g	67.9	12.5	48.1	36.1	63.8	34.7	43.8

ing a gain of +14.7 points over the base model. Among unlabeled methods, **CCAE** achieves the best average performance, outperforming INTUITOR and TTRL by +4.7 and +2.4 points, respectively. These results show that **CCAE** is not tied to the Qwen3 backbone and remains effective on a model from a different family.

5.5 Ablation Study

Table 5 reports ablations on MATH-Train and DAPO-14k-MATH. Removing any single factor consistently degrades the average performance, confirming the three components make complementary contributions.

Intrinsic evidence ϕ_i is the dominant driver.

Removing ϕ_i reduces the average Pass@1 by -8.8 points on MATH-Train ($52.2 \rightarrow 43.4$) and -12.1 points on DAPO-14k-MATH ($52.1 \rightarrow 40.0$), with the latter approaching the base-model level. Without intrinsic gating, the estimator treats all rollouts as equally informative and becomes vulnerable to noisy trajectories, especially in harder regimes.

Extrinsic agreement ψ_i matters more on harder data. Removing ψ_i drops the score by -4.4

points on DAPO-14k-MATH but only -1.3 on MATH-Train. This asymmetric pattern indicates that agreement information becomes more important in harder regimes, whereas intrinsic evidence is often sufficient on the cleaner MATH-Train distribution.

Adaptive reliability r_g is critical for safe consensus use. Fixing $r_g \equiv 1$ retains v_i but applies it uniformly. This degrades DAPO-14k-MATH by -8.3 points ($52.1 \rightarrow 43.8$) *worse* than removing agreement rectification altogether. Unreliable consensus is not merely uninformative but actively harmful when applied indiscriminately; r_g is what makes consensus safe to use.

The gains of **CCAE** therefore do not stem from a naive combination of self-certainty and agreement, but from a principled decomposition into intrinsic evidence, extrinsic agreement, and adaptive reliability.

6 Conclusion

We introduced **CCAE**, a verifier-free and parameter-free advantage estimator for label-free RL of LLM reasoning. Instead of relying on a sin-

gle pseudo-signal, **CCAE** formulates unlabeled RL as a multi-signal estimation problem and decomposes the advantage into a base self-certainty term, an intrinsic evidence factor, and an extrinsic agreement factor, calibrating policy updates by both internal and external evidence. Across mathematical and free-form natural reasoning settings, including a harder DAPO-14k-MATH regime and a non-Qwen backbone, **CCAE** consistently improves over strong unlabeled baselines and remains stable in regimes where prior methods collapse. These results suggest that the central challenge of label-free self-learning is not the absence of pseudo-signals, but their reliability.

Limitations

CCAE relies on answer extraction and clustering to estimate agreement, which may become noisier in highly open-ended domains. Our experiments focus on reasoning tasks and do not cover capabilities such as dialogue safety, tool use, or long-context instruction following. Finally, while **CCAE** itself requires no gold labels, our analysis of baseline collapse uses gold-label feedback to identify the best pre-collapse checkpoints of prior baselines, providing an optimistic comparison; developing fully label-free checkpoint selection remains an important future direction.

Acknowledgments

We used ChatGPT for language polishing and manuscript editing. All scientific content, analyses, and conclusions were reviewed and verified by the authors.

References

- Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*.
- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267.
- Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.
- Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.
- Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiad-bench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.
- Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*.
- Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn.

2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, et al. 2025. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Ilya Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.
- Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.
- Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. 2025. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.
- Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*.
- Yifan Zhang and Math-AI Team. 2025. [American invitational mathematics examination \(aime\) 2025](#).
- Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2025. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.
- Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. 2025. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*.

A Additional Method Details

A.1 Training Hyperparameters

We report the training hyperparameters used in all experiments in Table 6. Within each base model, all methods share the same training configuration to ensure a fair comparison. The default column applies to Qwen3-4B-Base and Qwen3-8B-Base, while OctoThinker Hybrid-8B-Base uses a longer completion length due to its different pretraining recipe and reasoning style. Only hyperparameters that affect the learned policy or evaluation are listed; unspecified fields follow the default settings of the verl framework.

A.2 Benchmark Details

We evaluate on eight benchmarks in total. The math and STEM benchmarks (**MATH-500**, **AIME25**, **AMC23**, **OlympiadBench**, **MMLU_STEM**, **GPQA-Diamond**) follow the standard evaluation protocols used in prior label-free RL work.

For evaluation after training on Natural Reasoning, we additionally use:

- **GPQA**: a graduate-level multiple-choice question-answering benchmark spanning biology, chemistry, and physics, written by domain experts.
- **MMLU-Pro**: an upgraded version of MMLU with more reasoning-heavy questions and a 10-way (rather than 4-way) answer set. We report accuracy on its four super-categories—STEM, Humanities, Social Sciences, and Other—along with the official overall average.

Table 6: Training hyperparameters used in all experiments.

Parameter	Default	OctoThinker Hybrid-8B-Base
Learning Rate	3×10^{-6}	3×10^{-6}
Epochs	1	1
Batch Size	128	128
Group Size	16	16
Rollouts per Prompt	16	16
KL Penalty (β)	0.001	0.001
Max Prompt Length	512	512
Max Completion Length	4096	8192
Temperature	1.0	1.0
LR Scheduler	Cosine	Cosine
Warmup Ratio	0.1	0.1
Optimizer	AdamW	AdamW

For all benchmarks, we sample 16 completions under temperature 0.6 and top- p 0.95 and report the average performance.

A.3 Baseline Details

- **INTUITOR** (Zhao et al., 2025) uses self-certainty as an intrinsic reward to train reasoning models without gold labels. The reward is derived from the model’s own confidence and optimized using a group-based RL objective.
- **TTRL** (Zuo et al., 2025) constructs pseudo target answers via majority voting over sampled rollouts; these pseudo targets are then reinforced during policy optimization. Since exact symbolic matching is not applicable to free-form answers, we additionally implement **Pseudo-TTRL** for the natural reasoning setting, where rollouts are clustered by answer equivalence and the largest cluster is used as the pseudo label.
- **EMPO** (Zhang et al., 2025) is a fully unsupervised reasoning method that constructs rewards by minimizing semantic entropy across multiple sampled solutions.

B Prompts and Model Completions

This section provides the prompt template used during training and evaluation. We use the same system prompt for all base models and all methods in our experiments.

B.1 Training Prompts

System prompt used in experiments.

You are a helpful AI Assistant, designed to provide well-reasoned and detailed responses. You **FIRST** think about the reasoning process step by step and then provide the user with the answer. Please enclose your final answer in the box: `\boxed{ Your Answer }`. Please stop generation immediately after outputting the box.

C Algorithm and Implementation Details

C.1 Answer Extraction and Semantic Clustering

For the Natural Reasoning setting, **CCA**E, **Pseudo-TTRL**, and **EMPO** use the same answer-extraction and semantic-clustering pipeline to ensure a controlled comparison. Each response is prompted to place its final prediction inside `\boxed{ }`, and we extract only the content of the final `\boxed{ }` while excluding the preceding reasoning trace from clustering.

Following the free-form clustering protocol used by **EMPO** (Zhang et al., 2025), we apply the same semantic-equivalence classifier to pairs of extracted answers. Two answers are considered equivalent only when they are judged to imply each other in both directions. The resulting pairwise equivalence relations are converted into semantic answer clusters, and the empirical probability of each cluster is estimated by its fraction of rollouts in the group.

The same extraction and clustering procedure is shared across methods. Pseudo-TTRL uses the largest cluster as its pseudo-label, EMPO uses cluster probability to construct its semantic reward, and **CCAE** uses the same cluster assignments to compute answer-level agreement. Therefore, these methods differ in how the resulting cluster statistics are converted into training signals rather than in answer parsing or semantic clustering.

Importantly, the semantic-equivalence classifier is not provided with gold answers and does not evaluate correctness. It only determines whether two generated final answers express the same meaning, and therefore serves as an equivalence classifier rather than a correctness verifier.

C.2 Algorithm Details of CCAE

This section details the **CCAE** advantage rectification procedure for a single prompt group in algorithmic form. The procedure takes as input the group of sampled rollouts, their internally computed self-certainty scores, and the extracted final answers, and returns rectified advantages that are subsequently consumed by the standard GRPO policy update. Importantly, the procedure is fully self-contained at the group level: it requires no cross-group statistics, no reward model, no learned verifier, and no additional trainable parameters. This makes **CCAE** a drop-in replacement for the advantage estimator in existing GRPO training pipeline.

Inputs and outputs. For each prompt x in a training batch, the policy π_θ samples a group of G rollouts $\{y_i\}_{i=1}^G$. For each rollout, we compute the per-token next-token distributions during generation and the corresponding self-certainty score s_i defined in Eq. (3). We also extract the final answer $\hat{a}_i$ from each rollout using the same answer parser as our voting-based baselines (TTRL and Pseudo-TTRL) to ensure fair comparison. The output of the procedure is the set of rectified advantages $\{\tilde{A}_i\}_{i=1}^G$, one scalar per rollout, which is then used by the clipped GRPO loss in Eq. (13).

Pseudocode. Algorithm 1 summarizes the full procedure. Lines 1–2 compute the base self-certainty advantage by within-group standardization. Lines 3–6 extract the agreement view: rollouts are clustered by answer equivalence, and the cluster proportions are used to compute the group-level reliability r_g . Line 7 computes the per-

rollout vote-direction score v_i . Lines 8–9 form the intrinsic evidence factor ϕ_i and the extrinsic agreement factor ψ_i , respectively. Line 10 multiplies the three terms to obtain the final advantage estimate.

Algorithm 1 CCAE: Consistency-Calibrated Advantage Estimation (per prompt group)

Require: Rollouts $\{y_i\}_{i=1}^G$ with self-certainty scores $\{s_i\}_{i=1}^G$ and extracted answers $\{\hat{a}_i\}_{i=1}^G$; small constant ϵ

- 1: Compute group statistics μ_s, σ_s over $\{s_i\}_{i=1}^G$
- 2: Standardize: $z_i \leftarrow (s_i - \mu_s)/(\sigma_s + \epsilon)$, set $A_i^{\text{base}} \leftarrow z_i$
- 3: Cluster rollouts by answer equivalence using $\{\hat{a}_i\}$
- 4: Compute cluster proportions $\{p_{a_i}\}$, dominant proportion $p_{\max}$, number of clusters K_g
- 5: Uniform baseline: $p_0 \leftarrow 1/K_g$
- 6: Group reliability: $r_g \leftarrow 1$ if $K_g = 1$, else $r_g \leftarrow (p_{\max} - p_0)/(1 - p_0 + \epsilon)$
- 7: Vote direction: $v_i \leftarrow \text{clip}((p_{a_i} - p_0)/(p_{\max} - p_0 + \epsilon), -1, 1)$
- 8: Intrinsic factor: $\phi_i \leftarrow |z_i| \cdot \mathbf{1}[z_i > 1 - r_g]$
- 9: Extrinsic factor: $\psi_i \leftarrow (1 - r_g) + r_g \cdot v_i$
- 10: Final advantage: $\tilde{A}_i \leftarrow A_i^{\text{base}} \cdot \phi_i \cdot \psi_i$
- 11: **return** $\{\tilde{A}_i\}_{i=1}^G$

Interpretation of the factor decomposition.

The three factors play distinct but interacting roles. The base advantage A_i^{base} encodes the relative confidence of rollout i within its group, but is not directionally reliable on its own. The intrinsic factor ϕ_i acts as a salience gate: only rollouts whose normalized self-certainty z_i exceeds the threshold $1 - r_g$ are allowed to contribute to the update. When the group exhibits strong answer agreement (r_g large), the threshold relaxes and most confident rollouts pass the gate; when agreement is weak (r_g small), the gate tightens and only rollouts with strongly above-average self-certainty are retained. The extrinsic factor $\psi_i = (1 - r_g) + r_g \cdot v_i$ interpolates between two regimes: when consensus is unreliable, $\psi_i \approx 1$ and the estimator effectively reduces to a self-certainty-driven update; when consensus is concentrated, $\psi_i \approx v_i$ and the update direction is calibrated by majority alignment. This adaptive behavior is what allows **CCAE** to remain stable across both clean and noisy reasoning regimes.

Edge cases. We handle several edge cases explicitly to ensure numerical stability. (i) If $\sigma_s = 0$ (all rollouts in the group share the same self-certainty), the standardization step in Line 2 yields $z_i = 0$ for all i , and the final advantage is zero. This prompt contributes no gradient, which is the desired behavior when the intrinsic signal carries no within-group information. (ii) If $K_g = 1$ (all rollouts produce the same answer), we set $r_g = 1$ by convention and $v_i = 1$ for all i , so $\psi_i = 1$ and the update is driven purely by the gated intrinsic signal. (iii) For answer extraction failures (e.g., malformed outputs), the corresponding rollout is assigned a unique singleton cluster, which lowers its vote support and naturally down-weights its contribution.

Computational overhead. CCAE adds only $O(G)$ work per prompt group on top of standard GRPO: a single pass over self-certainty scores for standardization, a hash-based clustering over the G extracted answers, and a constant number of arithmetic operations per rollout. Self-certainty itself is computed as a byproduct of the rollout forward pass and requires no additional model calls. In our implementation, the wall-clock overhead of CCAE relative to vanilla GRPO with self-certainty rewards is below 1% across all training runs, making it effectively free at scale.

Implementation notes. We implement CCAE on top of the verl framework. The advantage estimator is inserted as a post-processing step after rollout collection and before the policy loss computation. All three factors A_i^{base} , ϕ_i , and ψ_i are computed under `torch.no_grad()` because GRPO treats the advantage as a stop-gradient scalar; only the importance ratio $\rho_i(\theta)$ in Eq. (13) carries gradients to the policy parameters. We use the same answer extraction and equivalence rules across CCAE, TTRL, and Pseudo-TTRL to guarantee that any observed performance gap reflects the advantage estimator rather than differences in answer parsing. The small constant ϵ in Lines 2, 6, and 7 is set to 10^{-8} throughout all experiments and is not tuned.