ANSWER IS NOT ENOUGH: CAUSAL MEDIATION FOR REASONING CREDIT ASSIGNMENT IN GRPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Group Relative Policy Optimization (GRPO) optimizes reasoning policies using verifiable final answer rewards. However, final answer correctness alone may not provide an adequate learning signal for the reasoning process. Our preliminary investigation shows that, when the answer condition is fixed, changing the reasoning process leads to an accuracy difference of up to 11.8%, indicating that the reasoning process provides information beyond the final answer. Motivated by this observation, we propose **CaMeL** (Causal Mediation Lens), a reasoning credit assignment method based on causal mediation that models the reasoning process as a mediator to estimate its contribution to the answer decision. CaMeL estimates this contribution with two signals. **Reasoning Causal Effect (RCE)** measures the effect of the reasoning process on the correct answer probability relative to direct answering, capturing the direction and magnitude of its contribution. **Reasoning Causal Specificity (RCS)** measures whether the change in the answer distribution induced by the reasoning process exceeds those induced by shuffled and irrelevant reasoning processes, retaining only contributions beyond the control reference. The two signals form a mediator contribution, which is incorporated into rollout scoring by adjusting the Reason credit without altering the final answer reward. Experiments show that CaMeL outperforms Vanilla GRPO on logical reasoning benchmarks and generalizes to out-of-distribution (OOD) settings. Quantitative analyses and human evaluation further show that the mediator contribution aligns well with human judgments of the reasoning process, with over 90% directional agreement among activated signals.

1 INTRODUCTION

With reinforcement learning with verifiable rewards (RLVR), large language models (LLMs) can further improve their capabilities in question answering (Brown et al., 2020), code generation (Chen et al., 2021), and mathematical reasoning (Lewkowycz et al., 2022). A representative method, Group Relative Policy Optimization (GRPO), uses final answer rewards from multiple sampled rollouts to optimize the policy (Shao et al., 2024). Consequently, each rollout provides the same outcome-derived signal throughout its reasoning process, raising a natural question: can final-answer correctness alone provide an adequate learning signal for the reasoning process?

Notably, the reasoning process and final answer correctness are not always aligned. As shown in Figure 1, case (b) reaches the correct answer despite weak reasoning, whereas case (c) contains strong reasoning but produces an incorrect answer. This mismatch exposes two limitations in GRPO: (1) **limited reward discrimination**, as reasoning processes with the same final answer outcome receive the same reward; and (2) **reward misalignment**, as potentially useful reasoning processes may still receive unfavorable rewards when paired with an incorrect final answer.

Our preliminary investigations quantify these limitations: (1) 68.1% of rollouts receive zero group-relative advantage, indicating that final answer rewards often fail to distinguish reasoning processes within the same problem. (2) Holding the answer condition fixed, replacing supportive with contradictory reasoning reduces accuracy by 9.4%–11.8%, showing that the reasoning process affects learning beyond answer correctness. (3) The reasoning process benefits low-confidence inputs but harms high-confidence ones.

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.

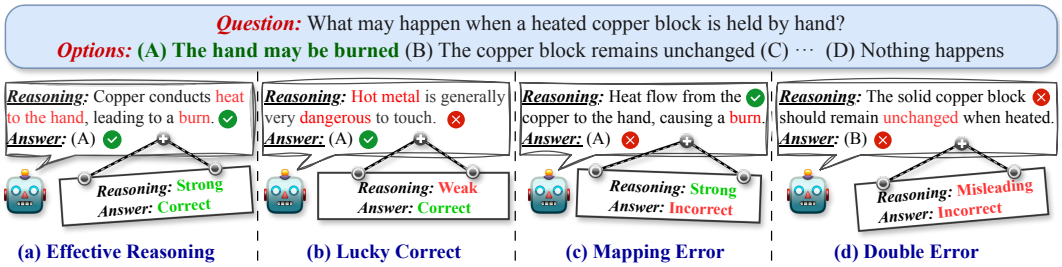

Figure 1: Schematic illustration of reasoning–answer mismatches.

These findings motivate us to estimate how the reasoning process contributes to the answer decision during RLVR. We therefore propose **CaMeL (Causal Mediation Lens)**, which treats the reasoning process as an intervenable mediator between the input and the answer decision. CaMeL combines **Reasoning Causal Effect (RCE)**, which quantifies the causal effect of the reasoning process on the correct answer probability relative to direct answering, with **Reasoning Causal Specificity (RCS)**, which evaluates the reliability of this effect using shuffled and irrelevant reasoning processes as controls. The resulting mediator contribution adjusts the Reason credit without altering the final answer reward, with confidence probing restricting mediator estimation to low-confidence inputs.

Through experiments, we obtain three findings. (1) CaMeL outperforms Vanilla GRPO across logical reasoning settings, with gains extending to out-of-distribution (OOD) and event relation reasoning settings. (2) The mediator contribution remains informative when GRPO provides no within-group discrimination and can adjust the reward-derived signal. (3) Human evaluation shows over 90% directional agreement between human judgments and activated mediator signals regarding the effect of the reasoning process on the answer decision. Our contributions are summarized as follows:

- We identify two limitations of GRPO: **limited reward discrimination** and **reward misalignment**, and design preliminary investigations to quantify their effects.
- We propose **CaMeL** to estimate the contribution of the reasoning process through **RCE** and **RCS** and use the resulting mediator contribution to adjust the Reason credit.
- We evaluate CaMeL across multiple models and tasks, demonstrating consistent improvements in reasoning performance and alignment between mediator signals and human evaluation.

2 Preliminary Investigations

To investigate the two limitations of **limited reward discrimination** and **reward misalignment**, we conduct three preliminary investigations to assess (1) whether final answer rewards distinguish rollouts, (2) how misalignment between the reasoning process and answer affects learning, and (3) whether the reasoning process is needed across different instances.

2.1 Do Final Answer Rewards Discriminate Among Rollouts in GRPO?

To quantify the **limited reward discrimination** of final answer rewards, we analyze 800 GRPO rollouts from Qwen3-1.7B (Yang et al., 2025) on ReClor (Yu et al., 2020). In GRPO, group-relative advantage is derived from group rewards and determines the optimization signal for each rollout.

As shown in Table 1, 68.1% of rollouts receive zero group-relative advantage, showing that final answer rewards fail to differentiate rollouts sharing the same answer.

Table 1: Distribution of group-relative advantages on ReClor.

Advantage	Count	Ratio (%)
Positive	145	18.1
Negative	110	13.8
Zero	545	**68.1**

2.2 What Is the Impact of Reward Misalignment?

To quantify the impact of **reward misalignment**, we conduct a controlled experiment on 1,000 ReClor examples. For each example, we construct reasoning processes P^+ and P^- that support or contradict the correct answer and pair them with correct A^+ or incorrect A^- answers.

This yields four conditions: P^+A^+, P^-A^+, P^+A^-, and P^-A^-. We fine-tune Qwen3-1.7B under each condition and evaluate accuracy and Positive Reason Preference (PRP), which measures preference for P^+ over P^-. Results are shown in Figure 2; details are provided in Appendix D.1.

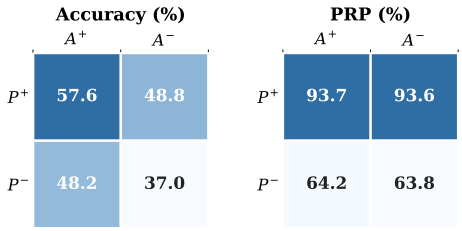

Figure 2: Effects of reasoning and answer conditions on ReClor.

(1) **Reasoning effect**: When the answer is fixed, replacing P^+ with P^- reduces accuracy and PRP, indicating that the reasoning condition affects both task performance and preference for reasoning processes supporting the correct answer.

(2) **Answer effect**: Starting from P^+A^+, replacing A^+ with A^- reduces accuracy from 57.6% to 48.8%, while further replacing P^+ with P^- reduces it to 37.0%. These results show that the answer and reasoning conditions have separable effects on task performance, indicating that the final answer alone does not fully characterize the learning contribution of the reasoning process.

2.3 When Is the Reasoning Process Needed?

Given that the reasoning process provides information beyond final answer rewards, we examine when it is needed. Using Qwen3-1.7B on ReClor, we compare it with direct answering across confidence levels. For each example, we define the margin confidence under direct answering as:

$$c_0 = p_0^{(1)} - p_0^{(2)}$$

where $p_0^{(1)}$ and $p_0^{(2)}$ denote the highest and second-highest candidate probabilities, respectively. We divide the examples into five equally sized groups according to c_0; details are provided in Appendix D.2.

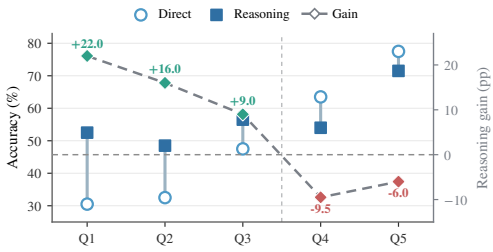

Figure 3: Confidence-dependent reasoning gain on ReClor.

As shown in Figure 3, the reasoning process improves accuracy in the three lower-confidence groups but reduces it in the two highest-confidence groups. This shows that the reasoning process is not always needed and may introduce errors when the model is confident under direct answering. We calibrate the confidence threshold at the boundary where reasoning gain becomes non-positive.

3 Method

Problem Formulation. Given a problem X with candidate answer set $\mathcal{A} = \{a_1, \ldots, a_K\}$, $y^+ \in \mathcal{A}$ denotes the correct answer. The policy model π_θ samples a group of rollouts $\{\tau_i\}_{i=1}^G$, where each rollout $\tau_i = (R_i, Y_i)$ consists of a reasoning process R_i and an answer $Y_i \in \mathcal{A}$. Each rollout receives a final answer reward based solely on Y_i, while CaMeL estimates the contribution of R_i to the answer decision and incorporates it into rollout scoring.

Framework Overview. To address the limitations of final answer rewards, we propose CaMeL, a three-stage framework. As illustrated in Figure 4, (1) **Confidence Probing** routes inputs to direct answering or the reasoning process; (2) **Causal Mediation Estimation** estimates the mediator contribution through RCE and RCS; and (3) **Asymmetric Credit Integration** adjusts the Reason credit using this contribution without altering the final answer reward.

3.1 Confidence Probing

Following Section 2.3, CaMeL uses the margin confidence c_0 under direct answering to determine whether a reasoning process is needed for each input. Specifically, c_0 is compared with a threshold τ, where τ is set at the empirical boundary at which the reasoning gain becomes non-positive:

$$\mathcal{X}_d = \{X \mid c_0(X) \geq \tau\}, \qquad \mathcal{X}_r = \{X \mid c_0(X) < \tau\}. \tag{1}$$

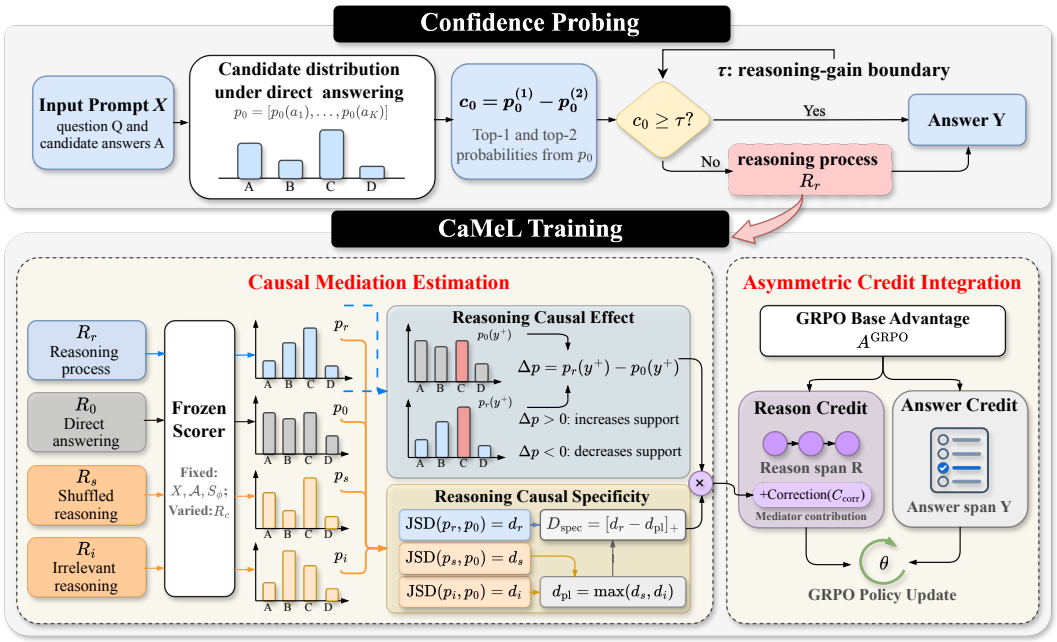

Figure 4: Overview of CaMeL.

For $X \in \mathcal{X}_d$, CaMeL uses direct answering, whereas for $X \in \mathcal{X}_r$, it invokes a reasoning process.

3.2 Causal Mediation Estimation

For mediation analysis, we treat the reasoning process R_r as an intervenable mediator between the input and the answer. Using a frozen scorer S_ϕ, we estimate the change in answer distribution induced by controlled interventions on R_r, holding X and the answer set $\mathcal{A}$ fixed. Accordingly, the estimated effect concerns the reasoning process and its influence on the scorer's answer decision.

3.2.1 Reasoning Causal Effect

The Reasoning Causal Effect **(RCE)** measures how R_r changes the scorer's support for the correct answer. We compare the direct answering baseline R_0 with R_r. Direct answering provides a reference in which no reasoning process is introduced. Relative to this reference, the change induced by R_r captures the effect of introducing the reasoning process on the answer decision. For $R_c \in \{R_0, R_r\}$, the frozen scorer produces a logit $z_\phi(a_k; X, R_c)$ for each candidate $a_k \in \mathcal{A}$. We normalize these logits over the candidate set:

$$p_c(a_k \mid X) = \frac{\exp z_\phi(a_k; X, R_c)}{\sum_{j=1}^{K} \exp z_\phi(a_j; X, R_c)}. \tag{2}$$

We estimate RCE as the change in the probability of the correct answer:

$$\Delta p = p_r(y^+ \mid X) - p_0(y^+ \mid X). \tag{3}$$

The sign of Δp indicates whether R_r moves the scorer's answer decision toward or away from the correct answer, while $|\Delta p|$ measures the magnitude of this change.

3.2.2 Reasoning Causal Specificity

RCE alone does not determine whether the observed change is specific to R_r, as comparable changes may also be induced by an order-shuffled version of R_r or an irrelevant reasoning process. We therefore introduce Reasoning Causal Specificity (**RCS**) to compare the change in the candidate distribution induced by R_r with those induced by the two control conditions R_s and R_{irr}. These controls are chosen to ablate two basic properties that distinguish decision-relevant reasoning from

generic text: R_s disrupts its organization while largely preserving the reasoning content, whereas R_{irr} removes task relevance while retaining an additional textual context. The construction of these control conditions is detailed in Appendix A.2. Unlike RCE, which focuses on the probability assigned to the correct answer, RCS considers the full candidate distribution.

Let

$$\mathbf{p}_c = [p_c(a_1 \mid X), \ldots, p_c(a_K \mid X)]\,. \tag{4}$$

For $c \in \{r, s, \text{irr}\}$, we measure the change in the candidate distribution relative to direct answering:

$$d_c = \text{JSD}_{\text{norm}}\left(\mathbf{p}_c, \mathbf{p}_0\right). \tag{5}$$

The normalized divergence and numerical stabilization are detailed in Appendix A.2. Hereafter, we use the larger of the two control-induced changes as the control reference, so that the effect of R_r is retained only when it exceeds both controls:

$$d_{\text{pl}} = \max\left(d_s, d_{\text{irr}}\right), \qquad D_{\text{spec}} = \text{clip}\left(d_r - d_{\text{pl}}, 0, 1\right). \tag{6}$$

Hence, $D_{\text{spec}} > 0$ only when R_r induces a greater change in the candidate distribution than both control conditions.

Finally, we combine Δp with D_{spec}:

$$C_{\text{corr}} = D_{\text{spec}}\,\Delta p. \tag{7}$$

Here, Δp captures the direction and magnitude of the change in correct answer support, while D_{spec} measures the effect retained relative to the controls. The resulting C_{corr} quantifies the mediator contribution for subsequent Reason credit adjustment.

3.3 Asymmetric Credit Integration

Since the final answer remains supervised by the final answer reward, CaMeL treats the Reason and Answer spans separately. Specifically, we retain GRPO's group-relative reward centering but omit within-group standardization, yielding $A^{\text{seg}}_{R,i}$ and $A^{\text{seg}}_{Y,i}$ for the two spans, respectively.

Based on this separation, CaMeL incorporates the mediator contribution only into the Reason credit:

$$A^{\text{CaMeL}}_{R,i} = A^{\text{seg}}_{R,i} + s_g\ \text{clip}\left(C_{\text{corr},i}, -\kappa, \kappa\right), \qquad A^{\text{CaMeL}}_{Y,i} = A^{\text{seg}}_{Y,i}, \tag{8}$$

where s_g controls the correction scale and κ bounds the magnitude of $C_{\text{corr},i}$ before scaling.

Thus, the mediator contribution modifies only $A^{\text{CaMeL}}_{R,i}$, while $A^{\text{CaMeL}}_{Y,i} = A^{\text{seg}}_{Y,i}$ remains unchanged. Both are applied to their corresponding spans in the GRPO policy update.

4 Experiments

4.1 Experimental Setup

Benchmarks and Metrics. We evaluate CaMeL on two logical reasoning benchmarks, **ReClor** (Yu et al., 2020) and **LogiQA2.0** (Liu et al., 2023). For OOD evaluation, we use four benchmarks: a social interaction benchmark **SocialIQa** (Sap et al., 2019), a physical commonsense benchmark **PIQA** (Bisk et al., 2020), an implicit reasoning benchmark **StrategyQA** (Geva et al., 2021), and an abductive reasoning benchmark α**-NLI** (Bhagavatula et al., 2019). To further evaluate CaMeL, we consider real-world event reasoning on the evidence-grounded abductive event reasoning benchmark **AER2026** (Cao et al., 2026). We report accuracy on the logical reasoning and OOD benchmarks, and use the official AER Score for AER2026, together with exact-set accuracy.

Models. We evaluate CaMeL with three backbones: **Qwen3-1.7B**, **Qwen3-4B-Instruct-2507** (Yang et al., 2025), and **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024).

Baselines. We report Base and SFT as reference settings and compare CaMeL with three RLVR baselines: **(1) Vanilla GRPO** (Shao et al., 2024), which derives group-relative advantages from final answer rewards; **(2) GRPO-λ** (Parthasarathi et al., 2025), which propagates reward information

Table 3: In-domain accuracy (%) on ReClor and LogiQA2.0 across three backbones. Avg. denotes the macro-average; best and second-best results are bold and underlined, respectively.

Method	Qwen3-1.7B		Qwen3-4B-Instruct-2507		Llama-3.1-8B-Instruct		Avg.
	ReClor	LogiQA2.0	ReClor	LogiQA2.0	ReClor	LogiQA2.0	
Base	57.67	47.88	74.00	62.10	48.00	38.00	54.61
SFT	58.00	48.41	74.00	61.46	53.67	46.07	56.94
Vanilla GRPO	58.67	50.53	80.67	66.67	65.00	50.32	61.98
OAR-G	57.67	49.47	81.33	67.30	65.00	52.87	62.27
GRPO-λ	58.00	49.68	82.00	68.15	63.67	50.85	62.06
CaMeL w/o Mediation	58.67	52.55	82.33	<u>68.47</u>	<u>65.67</u>	51.27	63.16
CaMeL w/o RCS	<u>59.67</u>	**52.76**	<u>82.67</u>	68.26	**66.00**	**55.20**	<u>64.09</u>
CaMeL w/o Asym. Integr.	59.33	51.80	82.33	67.30	<u>65.67</u>	52.65	63.18
CaMeL	**61.33**	<u>52.65</u>	**83.00**	**68.90**	**66.00**	<u>53.29</u>	**64.20**

through λ-returns without a learned critic; and **(3) OAR-G** (Li et al., 2026), which uses a gradient-based estimate of token influence on the final answer. To isolate the contribution of individual components, we consider three ablations: **w/o Mediation** removes mediator contribution while retaining confidence probing; **w/o RCS** removes RCS while retaining RCE; and **w/o Asymmetric Integration** applies the mediator contribution to both the Reason and Answer spans.

Implementation Details. All methods except Base start from the corresponding backbone without SFT initialization and use matched training data, rollout configurations, and optimization budgets. For CaMeL, S_ϕ is a frozen copy of the corresponding initialization model and remains fixed while π_θ is updated. The policy LoRA adapter is disabled during mediator estimation, requiring no additional learned scoring model. Key training and CaMeL-specific hyperparameters are summarized in Table 2. AER2026 uses the task-specific training setup described in Sec. 4.3.

Table 2: Main training configuration.

Configuration	**Value**
Optimizer	AdamW
Learning rate	1×10^{-5}
LR schedule	Linear, 5% warmup
Batch size	4
Optimization steps	400
Rollouts per prompt	8
LoRA r/α	32/32
KL β / clip ϵ	0.05/0.2
Reason / Answer weights w_R/w_Y	0.3/1.0
Mediator scale / clipping s_g/κ	0.5/0.2

Evaluation Protocol. The confidence threshold τ is calibrated on a disjoint calibration partition using the reasoning gain criterion in Sec. 3.1 and fixed thereafter. For OOD evaluation, the source checkpoint is evaluated on the target tasks without target training or model selection. AER2026 uses the task-specific evaluation protocol described in Sec. 4.3. Details are provided in Appendix A.3.

4.2 Main Results

Table 3 reports the main results on ReClor and LogiQA2.0 across the three model backbones. We make the following observations:

Overall effectiveness. CaMeL outperforms all RLVR baselines on the macro average. These gains indicate that incorporating the estimated mediator contribution provides complementary information beyond the final answer reward.

Across benchmarks. For each benchmark, we average accuracy across the three backbones. Relative to Vanilla GRPO, CaMeL improves by 2.00% on ReClor and 2.44% on LogiQA2.0. The gains on both benchmarks indicate that the improvement is not tied to the characteristics of a single logical reasoning dataset.

Across backbones. Across the three backbones, CaMeL outperforms all RLVR baselines when averaged over the two benchmarks, indicating that its effectiveness is not specific to a particular backbone. The largest gains over Vanilla GRPO occur on Qwen3-1.7B, suggesting that smaller backbones may be more sensitive to credit assignment quality and thus benefit more from CaMeL.

Table 4: Average OOD accuracy (%) over SocialIQa, PIQA, StrategyQA, and α-NLI.

Source Task	Method	Qwen3-1.7B	Qwen3-4B-Instruct-2507	Llama-3.1-8B	Avg.
ReClor	Base	54.93	71.19	54.28	60.13
	Vanilla GRPO	57.51	72.22	70.48	66.73
	OAR-G	59.34	72.41	73.11	68.28
	GRPO-λ	58.11	72.89	69.58	66.86
	CaMeL	**60.92**	**73.70**	**73.89**	**69.50**
LogiQA2.0	Base	54.93	71.19	54.28	60.13
	Vanilla GRPO	60.46	72.81	70.96	68.07
	OAR-G	59.42	73.06	72.08	68.19
	GRPO-λ	60.34	72.96	69.97	67.76
	CaMeL	**62.52**	**74.13**	**73.66**	**70.10**

Key components. Both **CaMeL w/o Mediation** and **CaMeL w/o Asym. Integr.** underperform CaMeL in all settings. This confirms that both the mediator contribution and its asymmetric application to Reason credit contribute to the overall performance.

Role of RCS. CaMeL w/o RCS achieves an average close to CaMeL, with mixed changes across settings. This reflects the filtering role of RCS: it suppresses effects that do not exceed the control reference, retaining reliable estimates of reasoning contribution rather than improving accuracy.

4.3 OOD Evaluation and Event Reasoning

We further evaluate CaMeL on OOD benchmarks and event reasoning.

OOD Evaluation. Table 4 reports average OOD accuracy over SocialIQa, PIQA, StrategyQA, and α-NLI. Detailed results are provided in Appendix B.1.

(1) CaMeL achieves the highest overall OOD accuracy, indicating that its improvements extend beyond the source logical reasoning benchmarks. (2) CaMeL trained on LogiQA2.0 achieves a higher absolute OOD average accuracy than when trained on ReClor. This may reflect the greater similarity in task format between LogiQA2.0 and the target tasks. (3) Stronger backbones do not always yield larger OOD improvements. This suggests that the gains from CaMeL are not simply attributable to stronger backbone performance.

Event Reasoning. We evaluate all methods on AER2026, with results shown in Table 5.

(1) CaMeL achieves the highest AER Score and exact-set accuracy, showing that its effectiveness extends to event reasoning.

(2) CaMeL does not outperform Vanilla GRPO on multi-answer cases, possibly because candidate-factorized training optimizes candidates independently, while multi-answer prediction requires joint correctness.

Table 5: AER2026 Results with Qwen3-1.7B.

Method	AER Score	Exact-set	Single	Multi
Base	14.13	12.75	9.40	27.68
SFT	68.22	66.67	69.60	53.57
Vanilla GRPO	69.04	67.16	69.80	**55.36**
OAR-G	70.18	68.14	71.20	54.46
GRPO-λ	68.14	66.34	69.00	54.46
CaMeL	**71.16**	**69.28**	**72.80**	53.57

5 Further Analysis

To better understand CaMeL, we examine (i) mediator–human evaluation alignment, (ii) its effect on Reason credit beyond group-relative advantage, and (iii) its variation with decision uncertainty, followed by a case study of the complete credit-assignment process.

5.1 Reasoning Contribution Analysis

To assess whether CaMeL's mediator signals capture the effect of the reasoning process on answer decisions, we conduct a blinded human evaluation of Qwen3-1.7B rollouts on ReClor and LogiQA2.0. Each reasoning process is evaluated as *Supports*, *Contradicts*, or *Insufficient Evidence*

relative to the correct answer. We assess the alignment of group-relative advantage, Δp, and C_{corr} with human evaluation labels using S–C AUROC, Cliff's δ, directional agreement, and coverage.

Table 6: Human evaluation of GRPO advantage and mediator signals. Dir. Agree is computed on non-zero signals; Coverage denotes their proportion.

Dataset	Signal	S–C AUROC ↑	Cliff's δ ↑	Dir. Agree (%)	Coverage (%)
ReClor	GRPO Adv.	0.519	0.038	53.8	29.9
	Δp	**0.877**	**0.755**	86.2	100.0
	C_{corr}	0.807	0.615	90.5	48.3
LogiQA2.0	GRPO Adv.	0.515	0.030	52.8	38.3
	Δp	**0.927**	**0.854**	88.3	100.0
	C_{corr}	0.850	0.699	92.3	55.3

(i) Directional alignment. Δp better discriminates *Supports* from *Contradicts* than group-relative advantage, achieving AUROCs of 0.877 and 0.927 on ReClor and LogiQA2.0, respectively, with high directional agreement. These results show that RCE better captures whether the reasoning process supports or contradicts the correct answer.

(ii) Control filtering. C_{corr} covers 48.3% and 55.3% of samples on ReClor and LogiQA2.0, respectively, with directional agreement 4.3% and 4.0% higher than Δp over all samples. This shows that D_{spec} filters out effects that do not sufficiently exceed the control reference, retaining more reliable estimates of reasoning contribution at lower coverage.

5.2 Reasoning Credit Assignment Analysis

To examine whether CaMeL provides Reason credit information beyond group-relative advantage, we analyze Qwen3-1.7B rollouts on ReClor across group-relative advantage conditions. The results are summarized in Table 7. Details are provided in Appendix C.4.

(i) Credit recovery. In all-correct and all-wrong groups, CaMeL produces a nonzero correction δ_R, recovering within-group credit differentiation absent from final answer rewards.

(ii) Credit adjustment. In mixed groups, when the group-relative advantage is nonzero, the activated mediator contribution can either counteract or reinforce the existing Reason credit, indicating that it adjusts Reason credit based on the final answer reward rather than simply adding extra credit.

Table 7: Reason-credit effects by GRPO group.

Adv.	**Group**	**Effect**	**Ratio**
Zero	All-correct	Recovery	41.25%
	All-wrong	Recovery	57.28%
Nonzero	Mixed	Counteract	30.00%
		Reinforce	37.78%

5.3 Decision Uncertainty Analysis

To examine how the effect of CaMeL varies with decision uncertainty, we divide 500 ReClor examples into low- and high-uncertainty groups using normalized entropy under direct answering. The results are shown in Table 8, with the entropy definition and split criterion provided in Appendix C.3.

(i) Low uncertainty. Although C_{corr} remains active for a subset of low-uncertainty examples, CaMeL slightly underperforms Vanilla GRPO in this group. This suggests that when the model is confident under direct answering, additional Reason credit adjustment provides limited benefit and may introduce unnecessary corrections.

Table 8: Accuracy and mediator activity.

Uncertainty	**Accuracy**		**Gain**	C_{corr} **Active**
	GRPO	CaMeL		
Low	**61.2%**	60.4%	−0.8 pp	31.6%
High	48.0%	**49.6%**	**+1.6** pp	**52.4%**

(ii) High uncertainty. In contrast, CaMeL outperforms Vanilla GRPO in the high-uncertainty group, where C_{corr} is more frequently activated. This suggests that Reason credit adjustment is more beneficial under greater decision uncertainty, supporting confidence probing.

5.4 Case Study

Figure 5 illustrates how CaMeL works on ReClor. **(a)** Vanilla GRPO predicts the correct answer C, but its reasoning process does not capture the decision-relevant logic. **(b)** Confidence probing determines that a reasoning process is needed, which substantially increases the scorer's support for C, yielding a positive RCE. **(c)** This shift exceeds those induced by shuffled and irrelevant reasoning processes, so RCS retains the effect and produces a positive C_{corr}. **(d)** The resulting mediator contribution reinforces the Reason credit. Overall, this example shows how CaMeL estimates the contribution of the reasoning process beyond answer correctness and incorporates it into Reason credit.

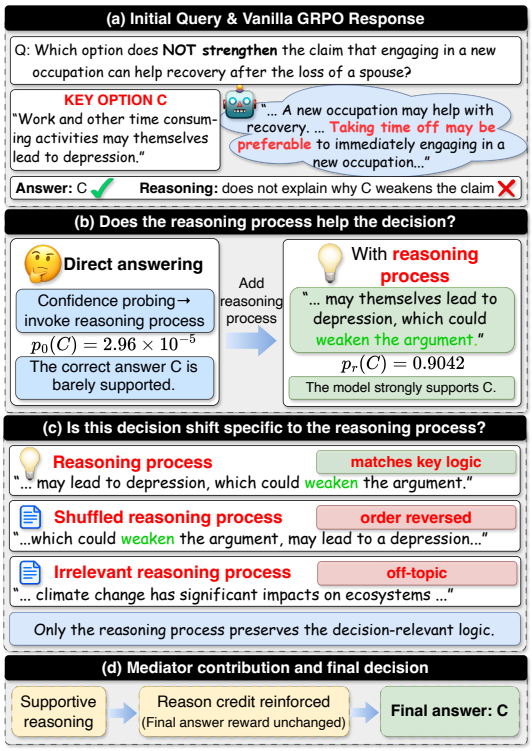

Figure 5: Case study of CaMeL on ReClor.

6 Related Work

6.1 LLM Reasoning

Chain-of-thought (CoT) prompting elicits intermediate reasoning steps to improve task performance (Wei et al., 2022). Reasoning can be enhanced at inference time via prompting (Kojima et al., 2022), path aggregation (Wang et al., 2023), structured search (Yao et al., 2023), and test-time scaling (Snell et al., 2025; Muennighoff et al., 2025), or through post-training (Yu et al., 2024), process supervision (Lightman et al., 2024; Wang et al., 2024), and verifiable rewards (Guo et al., 2025). Recent work further investigates how reasoning relates to model decisions using perturbations (Turpin et al., 2023; Lanham et al., 2023) and causal mediation (Paul et al., 2024).

Our work instead revisits credit assignment from final answer rewards in GRPO from the perspective of how the reasoning process causally contributes to answer decisions.

6.2 RLVR

RLVR optimizes reasoning policies with verifiable final answer rewards. GRPO computes group-relative advantages from sampled responses without a learned critic (Shao et al., 2024). Subsequent work refines reward attribution through segment-level evaluation (Guo et al., 2026) or token-level redistribution (Li et al., 2026). In contrast, process supervision provides denser intermediate feedback via implicit process rewards (Yuan et al., 2025) or reinforcement learning with process-level credit assignment (Cheng et al., 2026).

Rather than refining reward attribution or introducing process-level supervision, our work incorporates the estimated causal contribution of the reasoning process to answer decisions into GRPO.

7 Conclusion

This work revisits whether final answer rewards alone sufficiently credit the reasoning process that produces the answer. We show that the reasoning process and final answer need not be aligned. By treating the reasoning process as an intervenable mediator, CaMeL complements final answer rewards while preserving their supervision. Across logical reasoning, OOD, and event reasoning, it remains informative when group-relative advantage is zero, can counteract or reinforce existing Reason credit, and aligns strongly with human evaluation. Overall, CaMeL demonstrates that the contribution of the reasoning process can be estimated separately from the final answer and incorporated into GRPO without replacing verifiable rewards.

AI USE STATEMENT

Generative AI tools were used only as auxiliary tools in the research workflow, mainly for language polishing, literature organization, code debugging, and limited feedback on experimental implementation. All research questions, methodological designs, experimental decisions, analyses, and conclusions were developed and determined by the authors. AI-assisted code and text were manually reviewed and verified before use. In addition, Qwen3-4B-Instruct-2507 was used to generate paired reasoning traces for the controlled intervention experiment, as described in the experimental setup. Generative AI was not used to replace human annotation or to generate or modify the reported experimental results. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work evaluates language models on established reasoning benchmarks and model-generated reasoning processes, without using sensitive personal data. Human evaluation is limited to model-generated reasoning processes. Two evaluators independently assess each process, with a third adjudicator resolving disagreements. Evaluators are compensated at $10 per hour, above the local minimum wage, and consented to the use of their evaluations.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce our experiments in the main paper and appendix, including datasets, models, baselines, training configurations, evaluation protocols, and CaMeL implementation details. Code and additional materials are included in the anonymous supplementary material.

REFERENCES

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*, 2019.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Pengfei Cao, Mingxuan Yang, Yubo Chen, Chenlong Zhang, Mingxuan Liu, Kang Liu, and Jun Zhao. SemEval-2026 task 12: Abductive event reasoning: Towards real-world event causal inference for large language models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, pp. 3659–3672, 2026.

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. *arXiv preprint arXiv:2107.03374*, 2021.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. *Advances in Neural Information Processing Systems*, 38:131646–131671, 2026.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in RL for large language models. *Advances in Neural Information Processing Systems*, 38:114399–114431, 2026.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Ziheng Li, Liu Kang, Feng Xiao, Luxi Xing, Qingyi Si, Zhuoran Li, Weikang Gong, Deqing Yang, Yanghua Xiao, and Hongcheng Guo. Outcome-grounded advantage reshaping for fine-grained credit assignment in mathematical reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24681–24693, 2026.

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.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. LogiQA 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962, 2023.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. GRPO-λ: Credit assignment improves LLM reasoning. *arXiv preprint arXiv:2510.00194*, 2025.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, 2024.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQA: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

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

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

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

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap your own mathematical questions for large language models. In *International Conference on Learning Representations*, 2024.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. ReClor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*, 2020.

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

A Implementation and Reproduction Details

This appendix provides implementation and evaluation details required for reproduction but omitted from the main text. Unless otherwise specified, methods within the same model–dataset setting use the same initialization, training data, rollout configuration, optimization budget, and evaluation protocol reported in the main paper.

A.1 Baseline Implementation

All RLVR baselines are implemented within the same training framework and use the shared configuration reported in the main text unless otherwise specified.

Vanilla GRPO. We implement Vanilla GRPO following Shao et al. (2024), including group-wise reward centering and within-group standard-deviation normalization.

GRPO-λ. We reproduce GRPO-λ following Parthasarathi et al. (2025). We use $\lambda = 0.99$, $\gamma = 1.0$, `trace_style=both`, an advantage floor of -0.1, and a log-ratio clip of 20.0. All remaining settings follow the shared configuration in the main experiments.

OAR-G. We adapt OAR-G from the authors' official implementation.[1] We use the answer-mean outcome statistic, a token-importance quantile of 0.4, boost factor 2.0, embedding-noise standard deviation 0.001, suppression floor 0.1, and sum-preserving normalization over the Reason span.

[1] `https://github.com/XINGYUN-AI-LAB/OAR`

A.2 CaMeL Implementation Details

Confidence calibration. For each logical reasoning benchmark, the original development set is deterministically partitioned into disjoint calibration and checkpoint-selection subsets at a $40/60$ ratio. Examples are deterministically ordered by the SHA-256 hash of a fixed seed and the example identifier, and the first $\lfloor 0.4N \rfloor$ examples form the calibration subset. This yields 200 calibration and 300 selection examples for ReClor, and 627 calibration and 942 selection examples for LogiQA2.0.

The calibration subset is used only to determine the confidence threshold τ defined in Section 3.1. Within this subset, examples are ordered by the margin confidence c_0 under direct answering and divided into five equal-sized groups. Reasoning gain is computed within each group as the accuracy difference between using a reasoning process and direct answering. Scanning the groups from lower to higher confidence, τ is set at the empirical boundary where the reasoning gain becomes non-positive. The resulting threshold is fixed for the corresponding model–dataset setting throughout RL training and evaluation. The checkpoint-selection subset is not used for threshold calibration.

Mediator controls. For each generated reasoning process R_r, the direct-answering condition R_0 removes the reasoning content while keeping the problem X, candidate set $\mathcal{A}$, scoring prompt, and candidate ordering unchanged.

The shuffled control R_s is constructed within the same generated reasoning process rather than by permuting examples or rollouts. We first remove expressions that explicitly reveal the final answer and segment the remaining reasoning process into units using periods, question marks, exclamation marks, semicolons, or line breaks. The resulting units are placed in reverse order. If only one unit remains, its first whitespace-delimited token is moved to the end; sequences containing fewer than two tokens are left unchanged.

The irrelevant control R_{irr} uses the same fixed task-irrelevant placebo for all rollouts: *"This text is unrelated to the problem and supplies no evidence about any answer option."* It is not sampled from the training data, the current batch, or other rollouts. Across R_0, R_r, R_s, and R_{irr}, the problem, candidate set, scorer, scoring prompt, and candidate ordering remain fixed.

Mediator scoring. All mediator conditions are evaluated with the same frozen scorer S_ϕ. The policy parameters are not used to update S_ϕ during mediator estimation. For each condition, the scorer produces candidate logits $z_\phi(a_k; X, R_c)$ using the same registered candidate labels and candidate ordering, which are normalized over $\mathcal{A}$ as defined in Section 3.2. Scoring is deterministic and uses the same input construction across the four mediator conditions.

Numerical stabilization. Candidate probabilities are converted to float32 before divergence computation. Probabilities entering logarithms are lower-bounded by the float32 machine epsilon,

$$\epsilon_{\mathrm{num}} = \mathrm{finfo}(\mathrm{float32}).\mathrm{eps} \approx 1.19 \times 10^{-7}.$$

The Jensen–Shannon divergence is normalized by $\log 2$ and clipped to $[0, 1]$. The resulting D_{spec} is likewise clipped to $[0, 1]$.

Output parsing and span identification. The main experiments use an XML output contract with explicit `<reason>...</reason>` and `<answer>...</answer>` fields. A valid Answer span must resolve to a single registered candidate. The parser accepts normalized label forms such as `A`, `(A)`, and `Option A`, as well as an exact normalized match to one candidate text. Outputs without a uniquely recoverable candidate answer are treated as invalid; during evaluation, such outputs are counted as incorrect rather than discarded.

The Reason span is taken from the final valid `<reason>` field. A narrow structured recovery is permitted when imperfect or repeated answer markers still resolve unambiguously to the same candidate. This recovery does not search unrestricted free text and does not use the gold answer. Character-level spans returned by the parser are mapped to token masks through tokenizer-prefix decoding. If Reason and Answer token masks overlap, Answer tokens take precedence and are removed from the Reason mask.

Table 9: OOD accuracy (%) with ReClor as the source task. Best results are shown in bold.

Method	Qwen3-1.7B				Qwen3-4B-Instruct-2507				Llama-3.1-8B				Avg.
	SIQA	PIQA	SQA	α-NLI	SIQA	PIQA	SQA	α-NLI	SIQA	PIQA	SQA	α-NLI	
Base	56.96	53.59	52.84	56.33	66.84	75.52	**70.74**	71.67	45.91	55.33	53.28	62.60	60.13
Vanilla GRPO	59.67	57.02	53.28	60.05	67.81	79.00	69.87	72.19	69.09	77.26	59.39	76.17	66.73
OAR-G	60.59	58.60	54.15	**64.03**	68.53	78.62	69.43	73.04	**71.80**	**80.47**	63.32	**76.83**	68.28
GRPO-λ	60.80	57.34	53.28	61.03	**70.98**	80.52	66.81	73.24	68.99	79.92	55.90	73.50	66.86
CaMeL	**61.77**	**62.02**	**55.90**	63.97	70.42	**82.32**	68.56	**73.50**	70.21	77.09	**71.62**	76.63	**69.50**

A.3 Data Splits and Evaluation Details

Logical reasoning evaluation. Checkpoint selection is performed only on the checkpoint-selection partitions described above. Final evaluation uses deterministic decoding with one generation per example, temperature 0, top-$p = 1$, and a maximum generation length of 256 tokens. Reasoning process outputs follow the XML parsing procedure described in Appendix A.2; predictions without a valid candidate answer are counted as incorrect. For Qwen models, native thinking mode is disabled. Within each model–dataset setting, the same decoding and answer-extraction procedure is used across compared methods.

For inputs assigned to direct answering, no free-form reasoning process is generated. Instead, the registered candidate labels are scored directly under the model, and the candidate with the highest normalized likelihood is returned.

OOD evaluation. For each source benchmark, the checkpoint selected on the corresponding checkpoint-selection partition is evaluated unchanged on SocialIQa, PIQA, StrategyQA, and α-NLI. Target-task examples are not used for training, confidence calibration, or checkpoint selection. Decoding and answer extraction follow the same procedure used for the corresponding source-task model.

AER2026 evaluation. The AER2026 training, development, and test sets contain 1,819, 400, and 612 questions, corresponding to 7,276, 1,600, and 2,448 candidate-level episodes, respectively. Training runs for 200 optimization steps, with checkpoints evaluated at steps 50, 100, 150, and 200. Checkpoint selection uses development AER Score before the final test evaluation. At inference, candidate-level YES predictions are aggregated at the question level to form the final predicted relation set.

B Additional Experimental Results

B.1 Full OOD Results

Tables 9 and 10 report the per-target results underlying the averaged OOD results in Table 4.

Table 10: OOD accuracy (%) with LogiQA2.0 as the source task. Best results are shown in bold.

Method	Qwen3-1.7B				Qwen3-4B-Instruct-2507				Llama-3.1-8B				Avg.
	SIQA	PIQA	SQA	α-NLI	SIQA	PIQA	SQA	α-NLI	SIQA	PIQA	SQA	α-NLI	
Base	56.96	53.59	52.84	56.33	66.84	75.52	**70.74**	71.67	45.91	55.33	53.28	62.60	60.13
Vanilla GRPO	60.13	60.77	57.21	63.71	69.19	80.03	69.43	72.58	67.81	78.94	62.01	**75.07**	68.07
OAR-G	59.42	59.85	55.02	63.38	69.70	80.03	69.87	72.65	**70.88**	**81.34**	61.57	74.54	68.19
GRPO-λ	60.03	61.21	55.02	**65.08**	69.34	79.54	70.31	72.65	70.62	79.22	55.90	74.15	67.76
CaMeL	**61.00**	**64.69**	**59.39**	65.01	**71.95**	**82.32**	68.56	**73.69**	70.11	78.29	**71.62**	74.61	**70.10**

C Human Evaluation and Additional Analyses

C.1 Human Evaluation Protocol

We conduct human evaluation on reasoning processes generated by Qwen3-1.7B on ReClor and LogiQA2.0. Two evaluators independently assess each reasoning process, and disagreements are resolved by a third adjudicator. The evaluators are blinded to the group-relative advantage, Δp, D_{spec}, and C_{corr}. Each reasoning process is evaluated along two dimensions: reasoning contribution and reasoning correctness.

Reasoning contribution. **Supports** indicates that the reasoning process provides evidence or inference supporting the correct answer. **Contradicts** indicates that the reasoning process supports an incorrect alternative or contains reasoning inconsistent with the correct answer. **Insufficient Evidence** indicates that the reasoning process provides insufficient evidence either for or against the correct answer.

Reasoning correctness. **Correct** indicates that the key reasoning steps are valid. **Partially Correct** indicates that the reasoning process contains valid steps but is incomplete or contains non-critical errors. **Incorrect** indicates that the central reasoning is invalid. **Unverifiable** indicates that correctness cannot be determined from the available information.

Before adjudication, Cohen's κ is 0.906 for reasoning contribution and 0.951 for reasoning correctness.

Evaluation metrics. Let $\mathcal{R}$ denote the subset labeled *Supports* or *Contradicts*, and let $y_i \in \{+1, -1\}$ denote the corresponding human evaluation label. S–C AUROC measures discrimination between the two categories, while Cliff's δ measures the corresponding effect size.

For a signal s_i, directional agreement and coverage are defined as

$$\mathrm{Agree} = \frac{|\{i \in \mathcal{R} : s_i \neq 0,\, \mathrm{sign}(s_i) = y_i\}|}{|\{i \in \mathcal{R} : s_i \neq 0\}|},$$

and

$$\mathrm{Coverage} = \frac{|\{i \in \mathcal{R} : s_i \neq 0\}|}{|\mathcal{R}|}.$$

Directional agreement is therefore evaluated only on nonzero signals, whereas coverage measures the proportion of samples in $\mathcal{R}$ receiving a nonzero signal.

C.2 Additional Reasoning Contribution Statistics

After adjudication, the ReClor human evaluation contains 53 *Supports*, 13 *Insufficient Evidence*, and 34 *Contradicts* examples. The corresponding counts for LogiQA2.0 are 41, 6, and 53. The Supports–Contradicts subsets therefore contain 87 ReClor examples and 94 LogiQA2.0 examples.

Table 11 reports the distributions of Δp, D_{spec}, and C_{corr} across the three human evaluation categories. The Supports–Contradicts separation of Δp is significant on ReClor ($p = 3.40 \times 10^{-9}$) and LogiQA2.0 ($p = 1.58 \times 10^{-12}$). For C_{corr}, the corresponding Cliff's δ values are 0.615 and 0.699.

Table 12 reports coverage-aware directional statistics for the mediator signals.

Table 13 cross-tabulates reasoning contribution and reasoning correctness. On ReClor, all 47 reasoning processes evaluated as *Correct* are labeled *Supports*, whereas 33 of the 36 evaluated as *Incorrect* are labeled *Contradicts*. On LogiQA2.0, the corresponding counts are 36 of 36 and 49 of 51. Partially correct reasoning processes appear across all three contribution categories, showing that reasoning correctness and reasoning contribution capture related but distinct properties.

Table 11: Distribution of mediator quantities across human-evaluated reasoning contribution categories. Each dataset is reported in two rows: the median followed by [Q1, Q3].

Metric	Dataset	Supports	Insufficient	Contradicts
$\boldsymbol{\Delta p}$	ReClor	0.1297 [0.0078, 0.6138]	-0.0002 [-0.0020, 0.0479]	-0.0196 [-0.3522, -5.02×10^{-5}]
	LogiQA2.0	0.0784 [0.0015, 0.5836]	7.61×10^{-5} [-0.0007, 0.0740]	-0.0687 [-0.4606, -0.0004]
$\boldsymbol{D}_{\mathrm{spec}}$	ReClor	0.0000 [0.0000, 0.0912]	0.0000 [0.0000, 0.0120]	0.0227 [0.0000, 0.2066]
	LogiQA2.0	2.28×10^{-7} [0.0000, 0.1103]	0.0005 [0.0000, 0.0041]	0.0108 [0.0000, 0.1804]
$\boldsymbol{C}_{\mathrm{corr}}$	ReClor	0.0000 [0.0000, 0.0421]	0.0000 [-7.78×10^{-7}, 0.0000]	-6.69×10^{-11} [-0.0261, 0.0000]
	LogiQA2.0	0.0000 [0.0000, 0.0895]	0.0000 [-5.44×10^{-8}, 0.0000]	-3.51×10^{-6} [-0.0480, 0.0000]

Table 12: Coverage-aware directional analysis of the mediator correction. Coverage denotes the proportion of decision-relevant samples with $C_{\mathrm{corr}} \neq 0$. Covered agreement measures directional agreement conditioned on activation. Active and inactive agreement report the directional agreement of Δp within samples with non-zero and zero C_{corr}, respectively. All values are percentages.

Dataset	Coverage	Covered Agree.	Active Δp	Inactive Δp
ReClor	48.3	90.5	90.5	82.2
LogiQA2.0	55.3	92.3	92.3	83.3

C.3 Decision Uncertainty Details

Decision uncertainty in Section 5.3 is measured by the normalized entropy of the candidate distribution under direct answering:

$$H_{\mathrm{norm}}(p_0) = -\frac{1}{\log K}\sum_{k=1}^{K} p_0(a_k)\log p_0(a_k).$$

For the 500 ReClor examples used in this analysis, the low-/high-uncertainty split uses

$$H_{\mathrm{norm}} = 0.322709.$$

Mediator activation is defined as $C_{\mathrm{corr}} \neq 0$. The Spearman correlations between normalized entropy and D_{spec} and $|C_{\mathrm{corr}}|$ are 0.219 and 0.242, respectively.

C.4 Reasoning Credit Assignment Details

For the analysis in Section 5.2, rollout groups are categorized according to their final answer outcomes. An **all-correct** group contains only correct rollouts, an **all-wrong** group contains only incorrect rollouts, and a **mixed** group contains both.

Recovery denotes an active mediator correction $\delta_{R,i} \neq 0$ when $A^{\mathrm{seg}}_{R,i} = 0$. For $A^{\mathrm{seg}}_{R,i} \neq 0$, **Counteract** denotes

$$\mathrm{sign}(\delta_{R,i}) \neq \mathrm{sign}(A^{\mathrm{seg}}_{R,i}),$$

whereas **Reinforce** denotes matching signs.

The analysis starts from 800 rollouts collected in the dedicated Qwen3-1.7B/ReClor analysis run. We retain rollouts that enter the policy loss and have a valid tagged answer, a non-empty Reason span, no parser warning, and a valid routing key linking the rollout to its mediator statistics. This yields 786 valid rollouts. The remaining 14 contain malformed closing Answer tags, preventing valid answer extraction; consequently, they do not enter the policy loss and are excluded from this analysis.

Table 13: Relationship between reasoning correctness and human-evaluated contribution. Each entry reports the number of reasoning processes in the corresponding category.

Dataset	Reasoning Correctness	Supports	Insufficient	Contradicts
ReClor	Correct	47	0	0
	Partially correct	6	10	1
	Incorrect	0	3	33
LogiQA2.0	Correct	36	0	0
	Partially correct	5	3	4
	Incorrect	0	2	49
	Unverifiable	0	1	0

D Additional Details for Preliminary Investigations

D.1 Reasoning–Answer Intervention

Paired reasoning process construction. The controlled intervention in Section 2.2 uses 1,000 ReClor training examples. For each example, Qwen3-4B-Instruct-2507 jointly generates a paired positive and negative reasoning process, P^+ and P^-, in a single generation. P^+ is instructed to use sound reasoning that supports the full text of the gold option, whereas P^- contains a plausible but identifiable reasoning error and supports the full text of a predetermined incorrect option.

Both reasoning processes are required to be self-contained and comparable in length. They may not mention option labels, use phrases such as "the answer is," or provide a final answer. Each reasoning process is limited to 100 words and returned in separate `<reason_pos>` and `<reason_neg>` fields. Teacher generation is deterministic, using temperature 0, top-$p = 1$, one generation per example, and Qwen native thinking disabled. If the paired output violates the required structure, the example is regenerated once using the same deterministic decoding with a stricter formatting instruction.

Incorrect-answer selection. For each example, A^- is selected deterministically from the non-gold candidate labels. Let $\mathcal{A}^-$ denote the ordered set of non-gold labels. Using a fixed seed and the example identifier, we compute

$$h = \mathrm{SHA256}\left(\texttt{seed:example_id}\right),$$

convert the first eight digest bytes to an integer, and select

$$A^- = \mathcal{A}^-\left[h \bmod |\mathcal{A}^-|\right].$$

Thus, A^- is not selected by the teacher, model probability, or example difficulty.

The same P^+, P^-, A^+, and A^- are reused to construct all four intervention conditions,

$$P^+A^+, \qquad P^-A^+, \qquad P^+A^-, \qquad P^-A^-,$$

so comparisons differ only in the manipulated reasoning or answer condition.

Positive Reason Preference. For a reasoning process $P = (w_1, \ldots, w_T)$, we compute its length-normalized log-likelihood as

$$\ell(P) = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t \mid X, w_{<t}).$$

Positive Reason Preference (PRP) is then defined as

$$\mathrm{PRP} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\ell(P_i^+) > \ell(P_i^-)\right].$$

D.2 Additional Diagnostic Setup

Final answer reward discrimination. The 800 rollouts analyzed in Section 2.1 are collected from a Qwen3-1.7B/ReClor diagnostic run over 100 optimization steps, with eight rollouts sampled for each problem instance used at a step. Group-relative advantages are computed from the corresponding final answer rewards.

Confidence-dependent reasoning gain. The analysis in Section 2.3 uses 1,000 ReClor training examples with Qwen3-1.7B. Direct answering and reasoning process inference use the same model checkpoint, candidate set, and evaluation procedure; they differ only in whether the reasoning process is invoked. Examples are ranked by the margin confidence under direct answering

$$c_0 = p_0^{(1)} - p_0^{(2)}$$

and divided into five equal-sized groups. For each group, reasoning gain is defined as

$$\mathrm{Gain} = \mathrm{Acc}_{\mathrm{reason}} - \mathrm{Acc}_{\mathrm{direct}}.$$

Across all 1,000 examples, 12% are answered correctly under direct answering but incorrectly after invoking the reasoning process.