CLARITY: Reasoning Consistency Alone Can Teach Reinforced Experts

Anonymous ACL submission

Abstract

Training expert LLMs in domains with scarce data is difficult, often relying on multiple-choice questions (MCQs). However, standard outcome-based reinforcement learning (RL) on MCQs is risky. While it may improve accuracy, we observe it often degrades reasoning quality such as logical consistency. Existing solutions to supervise reasoning, such as large-scale Process Reward Models (PRMs), are prohibitively expensive. To address this, we propose CLARITY, a cost-effective RL framework that enhances reasoning quality using only a small, general-purpose LLM. CLARITY integrates a consistency-aware reward mechanism with a 2-stage refine-then-monitor training pipeline to enhance reasoning consistency, and a dynamic data reformulation strategy to to better exploit limited data. Experiments demonstrate that CLARITY improves response consistency by 16.5% and accuracy by 7.5% over baselines. Human evaluations further confirm holistic improvements in coherence and professionalism. Thus, CLARITY offers a generalizable solution that enables smaller models to effectively guide expert models by reasoning consistency.

1 Introduction

Reinforcement learning (RL) has recently become as a popular paradigm for enhancing the reasoning ability of large language models (LLMs) (Guo et al., 2025; Kimi et al., 2025; Qwen, 2025), yielding significant improvements in math (Feng et al., 2025; Yang et al., 2025) and code (Wang et al., 2025a; Fan et al., 2025). However, in other domains such as law and medicine, high-quality training data is often scarce. Available resources are typically limited to materials like professional qualification exams, where the predominant format is the multiple-choice question (MCQ).

Training using MCQ is usually effective, as their deterministic answers allow for clear outcome rewards, which avoids the potential instability of

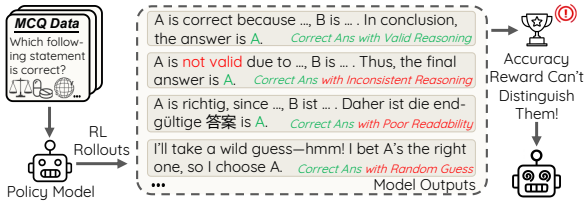

Figure 1: Illustration of risks in MCQ RL: rewarding only answer correctness neglects reasoning supervision, which may weaken reasoning quality during training.

training on open-ended tasks where a gold standard answer is sometimes ambiguous (GLM et al., 2025; He et al., 2025; Weng, 2024). However, it also leads to certain risks, as focusing solely on the final choice accuracy provides no guarantee of a faithful and high-quality reasoning process. As a result, the model may learn to find correct answers through flawed heuristics or even random guess, rather than acquiring strong reasoning abilities.

We conduct a pilot study training an LLM on a judicial examination MCQ dataset using standard outcome-based RL. As shown in Figure 2, while the final-answer accuracy significantly increases, the overall response quality progressively degrades. A typical manifestation of this quality decline is logical inconsistency. We find that the proportion of responses containing logical fallacies rises from 7% to 31%. This demonstrates that standard RL on MCQs not only fails to enhance reasoning quality, but even also produce unreliable models that generate incomprehensible reasoning.

Solving MCQ response quality issues is non-trivial. In domains like mathematics or code, the reasoning correctness can often be validated programmatically by external verifiers such as the Python interpreter (Lei et al., 2025; Zhao et al., 2025). For domains lacking such verifiers, current studies relies on Process Reward Models (PRMs) to supervise reasoning. This approach, however, typically requiring either the use of large LLMs like

GPT-4o as the PRM, or the fine-tuning of smaller LLMs on substantial high-quality, expert-annotated data (Chen et al., 2024a; Su et al., 2025). Both paths are resource-intensive, demanding significant budget or large expert-labeled corpora.

To overcome these challenges, we introduce **C**onsistency-aware **L**earning with Data-**A**ugmented 2-Stage **Rei**nforcement Strategy (CLARITY), a MCQ RL framework designed to enhance response quality using only a small, general-purpose LLM, without requiring any further fine-tuning on the domain-specific data.

CLARITY employs a consistency reward that assesses reasoning consistency within the response, penalizing incoherent outputs. And the reward is integrated into RL via a two-stage refine-then-monitor pipeline: Stage-1 refines the model's output to promote transparent, option-wise reasoning structure; Stage-2 relaxes format constraints and targets deeper reasoning, enhancing reasoning flexibility while preventing reward hacking. Additionally, to better exploit scarce data, CLARITY also apply a dynamic data reformulation strategy to improve data efficiency and overall training outcomes.

Our experiments validate the effectiveness of CLARITY, which achieves improvements of 16.5% in response consistency and 7.5% in reliable reasoning accuracy over standard RL baselines. Human evaluations further confirm that by focusing solely on the simple signal of logical consistency, CLARITY produces models with **holistic improvements** that are not only more coherent but also exhibit greater professionalism and readability. Moreover, by eliminating the need for large-scale teacher LLMs or expert-annotated datasets, CLARITY offers a cost-effective and generalizable solution across domains, enabling smaller, general-purpose LLMs to effectively guide the training of expert models to achieve both higher accuracy and superior reasoning quality.

Our contributions are summarized as follows: (1) We identify MCQ RL can yield superficially accurate yet unreliable, internally inconsistent reasoning. (2) We propose CLARITY, a novel MCQ RL framework that integrates consistency-aware mechanism with a refine-then-monitor pipeline to enhance reasoning consistency and a dynamic data reformulation strategy to maximize data utility. (3) We show CLARITY improves both accuracy and holistic reasoning quality, enabling smaller models to guide experts without requiring expert-level domain knowledge.

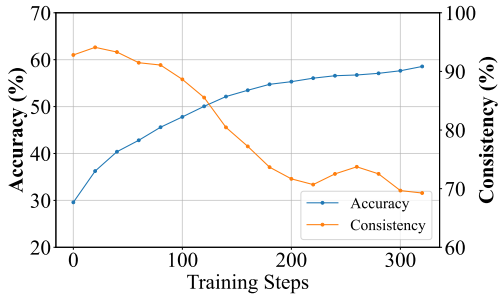

Figure 2: Response quality dynamics under GRPO training. The logical consistency declines over time.

2 Pilot Study: Response Quality Dynamic

we conduct a pilot study to investigate the response quality dynamics during MCQ RL training. We use the JEC-QA case-analysis dataset (Zhong et al., 2020) as the training data[1], which focuses on deep legal analysis and problem-solving skills through multi-choice questions. Throughout the training process, we assess the quality using DeepSeek-V3.

As shown in Figure 2, while the accuracy on final answer increases, we observe the overall response quality sharply degrades during training. Among those low-quality responses, a typical mode is logical inconsistency between the thinking process and the final answer, which we find more than 30% of the model-generated responses exhibited such contradictions after training.

These findings highlight that MCQ RL fails to enhance reasoning, producing models that reason unreliably and inconsistently. Given limited MCQ data resources without expert-annotation corpora or large PRM, we explore efficient strategies to address these challenges.

3 Methodology

We propose **C**onsistency-aware **L**earning with Data-**A**ugmented 2-Stage **Rei**nforcement Strategy (CLARity) to address response quality issues in MCQ RL. As shown in Figure 3, our approach integrates a consistency-aware learning mechanism into the data-augmented refine-then-monitor reinforcement pipeline. CLARITY enables smaller, general-purpose LLMs to guide expert model training, improving both reasoning quality and accuracy. See the complete algorithm in Algorithm 1.

3.1 Consistency-Aware Learning Mechanism

Our pilot study shows that outcome-based RL on MCQs fails to improve reasoning, often producing

[1]For more dataset details, see Section 4.

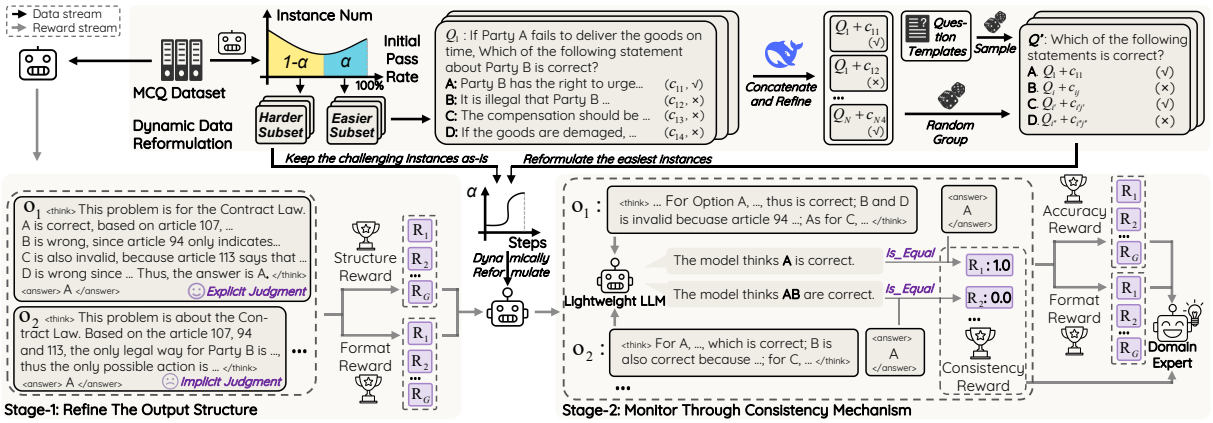

Figure 3: Overview of CLARITY, an efficient MCQ RL framework that trains expert models using only small, general-purpose LLMs. It combines a consistency mechanism for detecting inconsistencies, a refine-then-monitor training pipeline for improving reasoning quality, and a dynamic data reformulation for maximizing data utility.

responses with severe inconsistencies. Moreover, fully supervising the reasoning process requires either large commercial LLMs or costly expert-annotated corpora, which are often unavailable. Here, we explore another strategy to address the quality issue on MCQ RL, leveraging the consistency of model's judgments for each candidate option in its thinking trajectory as a reward signal. This is because a high-quality response must not only provide the correct final answer but also accurately judge the validity of each option, making consistency a useful proxy for reasoning reliability.

Specifically, our consistency-aware reward mechanism begins by separating the model's responses into two parts: the reasoning trajectory and the final answer. We then employ a separate, small general-purpose LLM as the consistency reward model, tasked with identifying the believed-correct options that the model endorses through its reasoning. A penalty is assigned if either the reward model fails to identify clear option judgments from the reasoning or the believed-correct options in the reasoning do not match the model's final answer. By penalizing such failures, our mechanism guides the model toward more coherent and reliable reasoning.

A key advantage of this mechanism is its minimal requirement for domain knowledge. The reward model only needs to comprehend basic correctness judgments (e.g., "*Option A is correct,*" "*B is invalid because ...*") within the response, which relies on fundamental semantic understanding, a capability already possessed by most general-purpose LLMs. Thus, expert-level models are unnecessary. Moreover, although our method focuses narrowly on logical coherence, we observe broader improvements in the holistic reasoning proficiency. See discussions in Section 5.5.

3.2 Data-Augmented Two-Stage Training

We introduce a two-stage refine-then-monitor pipeline to effectively integrate our consistency-aware reward into RL training. Additionally, to better exploit limited MCQ data, we propose a dynamic data reformulation strategy to improve data utilization and overall training outcomes.

2-Stage Refine-then-Monitor Training We integrate the proposed consistency reward into training process by a two-stage refine-then-monitor pipeline, which progressively improves consistency while preserving in-depth reasoning. In stage-1, we refine the model's output structure by encouraging option-wise reasoning, without considering the answer correctness. We apply regular expressions to identify and reward responses that analyze one option at a time, thus encouraging the model to explicitly state correctness judgments in reasoning. This ensures that responses remain LLM-friendly for accurate checking by reward models, and prevents reward hacking — if the model fails to present explicit judgments, the consistency reward model cannot reliably detect inconsistencies, which risks pushing the model toward overly superficial and simplified reasoning (Chen et al., 2025). Further discussion of this issue is provided in Section 5.4.

In stage 2, we monitor the model's responses using our proposed consistency reward model, and eliminate the requirement for response structure in stage-1 to avoid overfitting and encourage flexible reasoning. Additionally, we incorporate the answer

reward in stage-2 to optimize answer correctness. We design a strict reward mechanism that provides positive feedback only when the model selects exactly all the correct options and assigns zero reward in all other cases, which largely reduces the possibility of shortcut-based solutions during training and compels the model to develop a deeper reasoning. And a fixed format reward is applied in both stages, ensuring the model generates output within <think>...</think> and <answer>...</answer> tags.

Dynamic data reformulation Gradually removing easy instances and introducing harder instances during training can enhance model performance (An et al., 2025; Huang et al., 2025; Li et al., 2025b). Given the scarcity of high-quality training data, we employ a data augmentation method that does not require additional datasets. Instead, it refines too-easy data through deconstruction and reformulation to maximize data utility, curating more challenging and diverse data without the need for domain expertise.

Specifically, we first deconstruct each training instance into independent propositions by converting each candidate option into an atom statement through string concatenation of the original query statement and the option. The correctness of each proposition is determined by checking if the corresponding option is correct in the original problem. Next we refine these propositions using the LLM, polishing ill-formed phrasing for fluency and introducing fictional names and places to further diversify the data. See Appendix C.2 for data examples.

During training, based on the initial pass rate, we create new questions by dynamically reformulating the easiest α proportion of the original instances through randomly grouping, with answers derived from the correctness of each proposition and statements generated using predefined templates (e.g., *"Which of the following statements are correct?"*). The remaining $(1-\alpha)$ instances remain unchanged. By gradually increasing α, we progressively eliminate overly simple instances, transforming them into more challenging examples that promote better learning. This strategy offers an efficient and effective way to adjust the difficulty of training data and enhance training outcomes.

4 Experimental Setup

Datasets We focus specifically on two domains: law and medicine, where large-scale professional-level datasets exist solely in multiple-choice question (MCQ) format. We believe the observations from these areas are generalizable to other fields.

For legal reasoning, we use the JEC-QA case-analysis dataset (Zhong et al., 2020), which require deep legal analysis and reasoning in complex scenarios rather than rote memorization (Patterson, 1951). The dataset contains 10,561 case-analysis MCQs, split into 80% for training and 20% for validation. For medical reasoning, we use the MedQA-USMLE dataset (Jin et al., 2020), which evaluates a physician's ability to analyze medical concepts, apply principles and reason across multiple pieces of evidence. It contains 10,178 MCQs in the training set and 1,272 MCQs for validation.

Evaluation For in-domain validations, we report the ratio of correct and consistent instances (**Acc**$^+$ for abbreviation). This metric better reflects the model's true problem-solving ability, as it distinguishes genuine reasoning from superficial heuristics or random guessing. Additionally, we report the consistency ratio and the accuracy based only on the final generated answer (measured by exact match).

We also evaluate the model's generalizability on out-of-domain open-ended benchmarks covering different topics. For legal reasoning, we use LexEval (Li et al., 2025a) open-ended tasks, focusing on legal question answering and summary generation. For medical reasoning, we use PubMedQA (Jin et al., 2019), consisting of open-ended questions derived from biomedical research in PubMed. We employ Rouge-L as the evaluation metric for both benchmarks.

Implementation Details We use both Qwen-2.5-3B-Instruct and Qwen-2.5-7B-Instruct (Qwen et al., 2025) as the backbone model, resulting in CLARITY$_{3B}$ and CLARITY$_{7B}$ respectively. We use Qwen-2.5-7B-Instruct for the initial pass rate and the consistency reward calculation. For data augmentation and consistency evaluation, we utilize DeepSeek-V3 (DeepSeek-AI et al., 2024) and implement training with the GRPO (Shao et al., 2024) algorithm using the Verl framework.

In our two-stage refine-then-monitor training pipeline, we randomly sample 500 instances from the training dataset for stage-1, with the remaining data used for stage-2. When combining data reformulation with the pipeline, we apply it only in Stage-2, where we start with the original data and begin to reformulate at step 100 with $\alpha = 0.5$. See Appendix C.1 for more training details.

Method		Legal Reasoning			Medical Reasoning			Average			Non-MCQ	
		Acc^+	Cons%	Acc	Acc^+	Cons%	Acc	Acc^+	Cons%	Acc	Legal	Med
Vanilla Qwen	Qwen2.5$_{3B}$	17.5	74.3	19.3	34.7	77.7	38.2	26.1	76.0	28.8	17.7	17.4
	Qwen2.5$_{7B}$	28.5	92.3	29.3	42.5	88.7	49.4	35.5	90.5	39.4	24.9	17.0
RL Baselines	Standard RL	45.0	69.3	**58.5**	57.8	81.3	63.6	51.4	75.3	61.1	26.6	19.4
	PRM RL	42.0	68.2	54.7	60.2	**96.4**	60.8	51.1	82.3	57.8	25.5	22.2
Ours	Data Reformulation	47.4	75.3	56.9	62.4	89.0	65.3	55.1	82.2	61.1	25.3	**22.6**
	Refine-then-Monitor	52.3	93.8	56.3	59.7	89.3	62.3	56.0	91.6	59.3	23.6	18.3
	CLARITY$_{3B}$	33.9	**97.1**	34.2	50.1	86.0	55.1	42.0	91.6	44.7	22.4	19.1
	CLARITY$_{7B}$	**54.4**	94.1	56.4	**63.3**	89.5	**66.2**	**58.9**	**91.8**	**61.3**	**27.2**	**22.6**
	GPT-4o	39.5	96.6	39.9	83.0	99.2	86.3	61.3	97.9	63.1	27.5	22.2

Table 1: Performance on validation and non-MCQ benchmarks. Best scores are **bold**, with the second underlined.

Model	Legal		Medical		Avg.
	Iter	Shuf	Iter	Shuf	$\text{Acc}^+_{\text{Cons\%}}$
Qwen2.5$_{3B}$	$36.8_{66.7}$	$4.1_{47.1}$	$48.8_{79.9}$	$3.3_{43.2}$	$23.3_{59.2}$
Qwen2.5$_{7B}$	$54.5_{84.3}$	$14.5_{92.0}$	$67.8_{95.3}$	$6.9_{71.6}$	$35.9_{85.8}$
RL	$61.5_{87.1}$	$14.6_{55.4}$	$72.5_{94.7}$	$8.5_{61.8}$	$39.3_{74.8}$
PRM RL	$63.1_{88.4}$	$10.1_{40.4}$	$62.3_{97.2}$	$5.7_{83.9}$	$35.3_{77.5}$
Reform	$68.9_{91.3}$	$23.2_{78.5}$	$77.1_{95.8}$	$12.4_{86.0}$	$45.4_{87.9}$
Two-Stage	$70.5_{95.8}$	$23.7_{90.2}$	$75.1_{98.7}$	$12.7_{84.5}$	$45.5_{92.3}$
CLARITY$_{3B}$	$53.6_{86.8}$	$14.7_{97.3}$	$66.9_{95.0}$	$8.4_{85.3}$	$35.9_{91.1}$
CLARITY$_{7B}$	$\mathbf{70.5}_{93.3}$	$\mathbf{28.9}_{97.3}$	$\mathbf{81.0}_{99.4}$	$\mathbf{27.5}_{88.5}$	$\mathbf{52.0}_{94.6}$
GPT-4o	$60.0_{99.2}$	$20.3_{93.7}$	$88.0_{99.4}$	$40.4_{88.4}$	$52.2_{95.2}$

Table 2: Generalizability across formats.

5 Results and Analysis

5.1 Main Results

We report the main results in Table 1. Baselines, including standard RL and PRM RL, show degraded reasoning consistency across both domains, confirming the risks of applying RL to MCQs and the limitations of using small LLMs directly as PRMs. In contrast, **CLARITY achieves substantial gains in both consistency and accuracy**.

Moreover, **CLARITY shows strong generalizability**, with consistent gains on diverse unseen open-ended tasks. We further validate this by transforming the original validation set into two alternative formats with unchanged content: (1) Iterative Judgment (Iter.), where each option is judged independently, and (2) Shuffled MCQ (Shuf.), where options are recombined using the method in Section 3.2. Because options in the original MCQs are often correlated, so identifying one correct choice may suffice to solve the whole question. The new formats block such shortcuts to better test reasoning ability. As shown in Table 2, CLARITY maintains superior performance across both, which we attribute to our data reformulation strategy, as more diverse data usually leads to better generalization (Huang et al., 2025; Zhou et al., 2025).

Human evaluations confirm that **CLARITY improves holistic reasoning quality, yielding better professionalism and readability.** This shows that small general-purpose LLMs can effectively guide expert model training by monitoring consistency, producing models that even surpass large commercial systems like GPT-4o (OpenAI et al., 2024a).

5.2 How Does CLARITY Boost Reasoning?

In this section, we explore the underlying causes of the inconsistency, and identify the key reason for the improvement brought by the consistency-aware learning paradigm in CLARITY framework.

How can inconsistency affect reasoning? To better understand the mechanisms behind inconsistency, we categorize the phenomenon into three types: OVER EXCLUSION occurs when the reasoning process excludes all available option, yet the final answer selects some options it has just dismissed, suggesting that the model is able to consistently rule out all incorrect options, but it struggles to identify the correct one. OVER SELECTION arises when the LLM identifies multiple options as plausible in its reasoning, but the final answer contains only a partial subset of them. This points to the model can successfully identify and validate all the correct options but fails to verbalize the exclusion of some wrong options. DISSOCIATED ANSWER represents where the reasoning process explicitly concludes some options are the correct choice, but the final output exist a different, unrelated option. This indicates a severe breakdown between the intermediate reasoning and the final answer generation, indicating that the process of excluding irrelevant options and identifying the an-

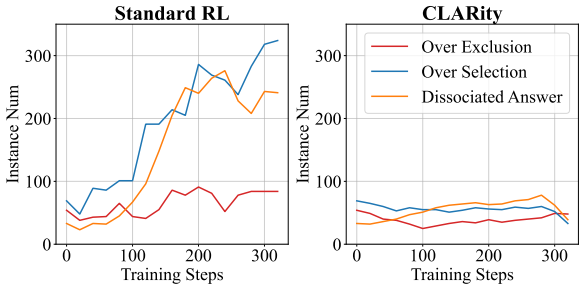

Figure 4: Training dynamics of three inconsistency types on Jec-QA validation set.

	Type	# Cons	# Incons	Cons Rate
Standard RL	Correct	951	285	76.9%
	Incorrect	513	365	58.4%
CLARITY	Correct	1150	42	96.5%
	Incorrect	840	82	91.1%

Table 3: Detailed statistics about consistency and correctness of different methods after RL training.

swers are both unreliable.

Vanilla RL training confuses the model's decision in excluding incorrect options, and causes hesitation on more harder tasks. Building upon our taxonomy, we now analyze how these inconsistencies evolve during training. We use DeepSeek-V3 as the annotator to track the outputs from the model undergoing standard RL training. As shown in Figure 4, the breakdown by type shows that the overall increase in inconsistencies is primarily driven by a significant rise in OVER SELECTION and DISSOCIATED ANSWER cases, while the OVER EXCLUSION case increase only marginally, suggesting that the primary failure is the model's growing reluctance to exclude wrong cases. Moreover, when examining consistency alongside accuracy (Table 3), we find that responses to incorrectly-answered questions exhibit much lower consistency than those correct ones. This suggests that standard RL makes the model more hesitant when facing difficult questions, failing to perform complex reasoning and thus offering very limited improvement in its actual reasoning capability.

CLARITY sharpens the model's capacity to exclude incorrect choices. As shown in Figure 4 and Table 3, by monitoring response consistency during training, all three types of inconsistency show a significant decrease compared to vanilla RL training, confirming our approach's effectiveness, especially in sharpening the model's capacity to exclude incorrect choices compared to the standard RL. We attribute these improvements to two key components: the consistency-aware learning mechanism (see discussion below) and the data reformulation pipeline (see Appendix D.1).

	Legal		Medical		Avg.
	Acc^+	Cons%	Acc^+	Cons%	$\text{Acc}^+_{\text{Cons\%}}$
Qwen2.5	28.5	92.3	42.5	88.7	$35.5_{90.5}$
Open-Ended Rewarding					
StrMatch RL	25.6	90.6	36.0	89.1	$30.8_{89.9}$
PRM RL	26.1	86.5	39.8	91.2	$33.0_{88.9}$
MCQ Rewarding					
StrMatch RL	45.0	69.3	57.8	81.3	$51.4_{75.3}$
PRM RL	42.0	68.2	60.2	**96.4**	$51.1_{82.3}$
CLARITY	**54.4**	**94.1**	**63.3**	89.5	$\mathbf{58.9_{91.8}}$

Table 4: Different reward modeling methods comparison. *StrMatch* denotes assigning a reward only when the generated answer exactly matches the ground truth.

5.3 Can Small Models Guide Experts?

Training expert-level LLMs typically requires large models like GPT-4 or additional training to augment domain knowledge (Chen et al., 2024a; Su et al., 2025). These methods are resource-intensive, requiring considerable financial investment or huge amount of expert-annotated data. We investigate whether smaller, general-purpose LLMs can effectively guide the training for expert model.

Small-scale LLMs fail as PRM. We first examine whether smaller open-source LLMs (Qwen2.5-7B-Instruct) can directly serve as the PRM, monitoring the reasoning process with the same instructions used for larger models. Responses are rewarded only if they are free from any possible error. As shown in Table 4, we observe two key points. First, treating questions as open-ended generation yields significantly lower accuracy, demonstrating that the non-deterministic nature of open-ended answers leads ineffective training. MCQs, with clear and deterministic answer, can provide more stable and generalizable improvements. Second, the 7B model fail to serve as an effective MCQ PRM, leading to results substantially inferior to CLARITY. This is likely due to a combination of limited domain expertise and inadequate model capacity, which together constrain its ability to accurately assess reasoning quality.

CLARITY enables guidance via monitoring consistency. It introduces a consistency-aware learn-

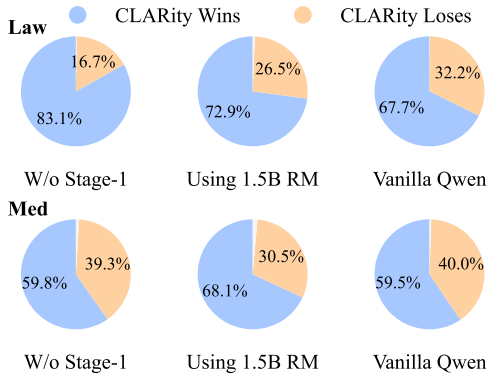

Figure 5: Reasoning quality comparison between CLARITY and different baselines: without Stage-1, using Qwen-1.5B as the consistency reward model, and the vanilla Qwen2.5-7B-Instruct.

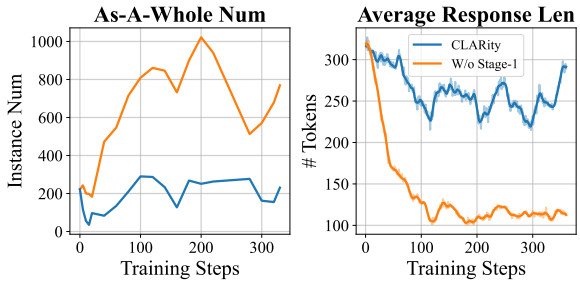

Figure 6: Training dynamics of *as-a-whole* response frequency and response length. In CLARITY, the first 15 steps correspond to Stage-1.

ing paradigm, allowing smaller LLMs to monitor response consistency in MCQs. As shown in Table 4, using only a 7B-scale LLM as its consistency checker, CLARITY significantly outperforms standard RL and PRM methods, and successfully cultivates an expert model with superior reasoning quality and accuracy. Further human evaluations in Section 5.5 confirm that CLARITY exhibit greater professionalism in their reasoning and are easier for readers to understand. These results demonstrate that small LLMs, when used to monitor consistency, can effectively curate expert LLMs.

We further investigate whether tiny-sized LLMs ($\leq$1.5B, deployable on consumer-grade GPUs) can also monitor reasoning consistency. First, we benchmark their inconsistency detection ability against DeepSeek-V3 as a proxy for ground truth[2]. The detection rates are substantially lower for tiny LLMs: 21% for Qwen2.5-0.5B and 40% for Qwen2.5-1.5B, compared to 94% for the 7B model. This indicates that the foundational semantic capabilities of these tiny models are insufficient even for the simpler consistency-checking task. Then we experiment with Qwen2.5-1.5B as the consistency reward model within the CLARITY framework, as shown in Figure 5, using tiny-sized reward models results in poorer response quality, as they encourage simpler, more superficial outputs that the weak consistency checker can easily parse, potentially leading to reward hacking.

5.4 How to Mitigate Reward Hacking?

Reward hacking may occur when tiny-sized LLMs are used as reward models, as their limited capacity prevents them from detecting inconsistencies. LLMs with relatively larger size, such as the 7B models, can identify inconsistencies, but they may still lack domain-specific knowledge, preventing them from distinguishing between substantive reasoning and superficially plausible shortcuts, which can also result in reward hacking. We propose a 2-stage refine-then-monitor pipeline in CLARITY to mitigate this, which we evaluate through experiments to demonstrate its necessity.

Refine-then-monitor pipeline offers flexible reasoning and avoids hacking. Our 2-stage refine-then-monitor pipeline first encourages explicit reasoning for each option and then uses a consistency reward model to monitor the responses. We categorize the responses into two types based on the observed patterns in model reasoning. *Explicit*: The reasoning includes an explicit and detailed analysis of options, typically appears in complex questions requiring diverse knowledge. *As-a-Whole*: The reasoning is generalized and presented without itemizing options, often uses in simpler questions requiring minimal thought.

As shown in Figure 6, in stage-1, the number of *As-a-Whole* responses decreases as we incentivize a transparent and structured format, making the reasoning process legible and easy for the consistency checker to evaluate accurately. In stage-2, we remove this structural constraint and monitor response quality with the consistency reward. We observe a gradual increase in *As-a-Whole* responses, which finally stabilizes. This demonstrates that the model has not merely memorized a format but has learned to flexibly apply the appropriate reasoning pattern based on task complexity, achieving consistently high-quality outputs.

Ablating the pipeline leads to reward hacking. To prove the necessity of our design, we ablated

[2]We validate the reliability of using DeepSeek-V3 as a proxy for ground truth in Appendix B.

the pipeline by removing the first stage, applying the reward LLM throughout the training. We assess response quality using LLM-as-a-Judge (Gu et al., 2025), evaluating answer rationality, professionalism, coherence, clarity, and ease of understanding.

As shown in Figures 5 and 6, removing the first stage leads to a significant degradation in response quality with much shorter responses. This is because the trained model quickly learns that simpler, more superficial reasoning is preferred. And it may even attempt to conceal complex reasoning to avoid penalties, resulting in reward hacking. This hypothesis is further supported by the statistic in Figure 6, where the number of *As-a-Whole* responses significantly increases in training, indicating that the model increasingly favor over-simplified reasoning, even when the questions demand substantial analysis. See Appendix E for case study and Appendix D.3 for further ablation on reward models.

5.5 Human Evaluation

We further demonstrate that CLARITY can improve the overall response quality beyond reasoning consistency through human evaluation.

Evaluation Metrics For both law and medicine, We recruit an expert who have passed profession examination to assess the quality of responses in 50 random samples from validation set. We establish three evaluation criteria focusing on different aspects, and can represent the overall quality when combined altogether: **Correctness** (Corr.), **Professionalism** (Prof.), **Readability** (Read.). We use a 5-point Likert scale, where 1 represents "very poor" and 5 represents "very good".

CLARITY improves professionalism and readability. The results in Table 5 show that CLARITY consistently achieves higher average scores than the vanilla RL model, demonstrating its effectiveness in enhancing overall reasoning quality by monitoring reasoning consistency. See Appendix B for more human evaluations.

	Standard RL			CLARITY		
	Corr.	Prof.	Read.	Corr.	Prof.	Read.
Legal	3.0	2.6	2.9	3.1	2.8	2.9
Medical	3.5	3.9	3.3	3.8	4.2	3.5
Average	3.3	3.3	3.1	**3.5**	**3.5**	**3.2**

Table 5: Human evaluation results on both domains.

6 Related Works

Reinforcement Learning for LLM Reasoning models, such as OpenAI o1 (OpenAI et al., 2024b) and DeepSeek-R1 (Guo et al., 2025), mainly focus on leveraging reinforcement learning methods like PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) to enhance LLM reasoning abilities. By rewarding accuracy and format, these methods achieve significant improvement across benchmarks in mathematics (Hu et al., 2025; Pan et al., 2025; Feng et al., 2025; Yang et al., 2025), logic (Xie et al., 2025), code generation (HuggingFace, 2025) and multi-hop QA (Song et al., 2025; Jin et al., 2025; Huang et al., 2025). However, few studies have adapt RL to other domains with only MCQ data available such as law and medicine (Chen et al., 2024a; Yu et al., 2025), where the clarity of the responses can't be guaranteed. We are the first to analyze the effectiveness of RL training in these domains from the perspective of response quality, focusing on logical consistency.

LLM Reasoning Consistency refers to whether the model's observed reasoning accurately reflects its internal thinking. Many studies highlights LLM consistency as a bottleneck in various settings, including incoherent logical reasoning (Kirchner et al., 2024; Ferreira et al., 2025; Arcuschin et al., 2025), resistance to verbalizing hints (Chen et al., 2025) or deliberately concealing its true capabilities (Meinke et al., 2025). To improve consistency, current research typically employs two ways: constructing high-quality supervised fine-tuning (SFT) data (Chen et al., 2024b; Paul et al., 2024; Ferreira et al., 2025; Chua et al., 2025), and filtering low-quality outputs during training (Wang et al., 2025b; Kirchner et al., 2024; Baker et al., 2025). Our work, focusing on improving data quality and dynamically rewarding high-quality responses, extends these approaches with the context of RL training.

7 Conclusion

We identify MCQ RL often produces superficially accurate yet inconsistent reasoning. We introduce CLARITY, an efficient RL framework that addresses this issue by guiding the expert model through consistency-aware refine-then-monitor training. Our method improves data utilization and delivers holistic enhancements in reasoning quality, including accuracy, professionalism, and readability, all without relying on large commercial models or expert-annotated corpora, thus offering a cost-effective and generalizable solution.

Limitations

Limited Exploration of Data Augmentation Approach Although this work demonstrates the effectiveness of our proposed data reformulation strategy, its exploration remains relatively limited. In particular, in this paper, we do not fully exploit its potential flexibility—for example, dynamically varying the number of candidate options per question, or mixing data from multiple distinct domains to create more diverse and challenging training samples. We believe these directions hold promise for further improving generalization and encourage future work to investigate them using our proposed data reformulation strategy further.

Training-Time Overhead In addition, while our CLARITY framework only relies on a relatively small 7B LLM as the consistency reward model, this still introduces substantial training-time overhead (from about 8 hours without reward modeling to ≥ 1 day with it). A more efficient alternative would be to replace the LLM reward model with a lightweight encoder such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019). However, this would require large-scale, high-quality labeled data to train these smaller models to perform reasoning comprehension and identify the model's believed-correct options—data that is currently difficult and costly to obtain. We therefore leave this as a promising direction for future research.

References

Chenxin An, Zhihui Xie, Xiaonan Li, Lei Li, Jun Zhang, Shansan Gong, Ming Zhong, Jingjing Xu, Xipeng Qiu, Mingxuan Wang, and Lingpeng Kong. 2025. Polaris: A post-training recipe for scaling reinforcement learning on advanced reasoning models.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. *Preprint*, arXiv:2503.08679.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. 2025. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *Preprint*, arXiv:2503.11926.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. Huatuogpt-o1, towards medical complex reasoning with llms. *Preprint*, arXiv:2412.18925.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Reasoning models don't always say what they think. *Preprint*, arXiv:2505.05410.

Yanda Chen, Chandan Singh, Xiaodong Liu, Simiao Zuo, Bin Yu, He He, and Jianfeng Gao. 2024b. Towards consistent natural-language explanations via explanation-consistency finetuning. *Preprint*, arXiv:2401.13986.

James Chua, Edward Rees, Hunar Batra, Samuel R. Bowman, Julian Michael, Ethan Perez, and Miles Turpin. 2025. Bias-augmented consistency training reduces biased reasoning in chain-of-thought. *Preprint*, arXiv:2403.05518.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 179 others. 2024. Deepseek-v3 technical report. *ArXiv*, abs/2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Lishui Fan, Yu Zhang, Mouxiang Chen, and Zhongxin Liu. 2025. Posterior-grpo: Rewarding reasoning processes in code generation. *Preprint*, arXiv:2508.05170.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. 2025. ReTool: Reinforcement Learning for Strategic Tool Use in LLMs. *Preprint*, arXiv:2504.11536.

Pedro Ferreira, Wilker Aziz, and Ivan Titov. 2025. Truthful or fabricated? using causal attribution to mitigate reward hacking in explanations. *Preprint*, arXiv:2504.05294.

GLM, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Bolei He, Xinran He, Run Shao, Shanfu Shu, Xianwei Xue, Mingquan Cheng, Haifeng Li, and Zhenhua Ling. 2025. Select to know: An internal-external knowledge self-selection framework for domain-specific question answering. *Preprint*, arXiv:2508.15213.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. 2025. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model. https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero.

Jerry Huang, Siddarth Madala, Risham Sidhu, Cheng Niu, Hao Peng, Julia Hockenmaier, and Tong Zhang. 2025. Rag-rl: Advancing retrieval-augmented generation via rl and curriculum learning. *Preprint*, arXiv:2503.12759.

HuggingFace. 2025. Open r1: A fully open reproduction of deepseek-r1.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *Preprint*, arXiv:2503.09516.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

Kimi, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. 2024. Prover-verifier games improve legibility of llm outputs. *Preprint*, arXiv:2407.13692.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Fangyu Lei, Jinxiang Meng, Yiming Huang, Tinghong Chen, Yun Zhang, Shizhu He, Jun Zhao, and Kang Liu. 2025. Reasoning-table: Exploring reinforcement learning for table reasoning. *Preprint*, arXiv:2506.01710.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2025a. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Advances in Neural Information Processing Systems*, 37:25061–25094.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025b. Limr: Less is more for rl scaling. *Preprint*, arXiv:2502.11886.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2025. Frontier models are capable of in-context scheming. *Preprint*, arXiv:2412.04984.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, and 243 others. 2024b. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. 2025. Tinyzero. https://github.com/Jiayi-Pan/TinyZero. Accessed: 2025-01-24.

Edwin W Patterson. 1951. Case method in american legal education: Its origins and objectives, the. *J. Legal Educ.*, 4:1.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *Preprint*, arXiv:2402.13950.

Qwen. 2025. Qwq-32b: The power of scaling rl.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and

24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297. ACM.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *Preprint*, arXiv:2503.05592.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *Preprint*, arXiv:2503.23829.

Sijie Wang, Quanjiang Guo, Kai Zhao, Yawei Zhang, Xin Li, Xiang Li, Siqi Li, Rui She, Shangshu Yu, and Wee Peng Tay. 2025a. Codeboost: Boosting code llms by squeezing knowledge from code snippets with rl. *Preprint*, arXiv:2508.05242.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2025b. Chain-of-probe: Examining the necessity and accuracy of cot step-by-step. *Preprint*, arXiv:2406.16144.

Lilian Weng. 2024. Reward hacking in reinforcement learning. *lilianweng.github.io*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning. *Preprint*, arXiv:2502.14768.

Ling Yang, Zhaochen Yu, Bin Cui, and Mengdi Wang. 2025. Reasonflux: Hierarchical llm reasoning via scaling thought templates. *Preprint*, arXiv:2502.06772.

Hongzhou Yu, Tianhao Cheng, Yingwen Wang, Wen He, Qing Wang, Ying Cheng, Yuejie Zhang, Rui Feng, and Xiaobo Zhang. 2025. Finemedlm-o1: Enhancing medical knowledge reasoning ability of llm from supervised fine-tuning to test-time training. *Preprint*, arXiv:2501.09213.

Yang Zhao, Chengxiao Dai, Wei Zhuo, Tan Chuan Fu, Yue Xiu, Dusit Niyato, Jonathan Z. Low, Eugene Ho Hong Zhuang, and Daren Zong Loong Tan. 2025. Agentict^{2}s:robust text-to-sparql via agentic collaborative reasoning over heterogeneous knowledge graphs for the circular economy. *Preprint*, arXiv:2508.01815.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jecqa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

Ruochen Zhou, Minrui Xu, Shiqi Chen, Junteng Liu, Yunqi Li, Xinxin Lin, Zhengyu Chen, and Junxian He. 2025. Does Learning Mathematical Problem-Solving Generalize to Broader Reasoning? *Preprint*, arXiv:2507.04391.

A Detailed Algorithm

The detailed algorithm of CLARITY is shown in Algorithm 1. N denote the total number of instances in the training dataset D, and M the number of candidate options for each multiple-choice question. Each problem instance $I_i = (Q_i, C_i, Ans_i) \in D$ consists of a query statement Q_i and M candidate options $C_i := \{C_{i1}, C_{i2}, \ldots, C_{iM}\}$, with an answer $A_i \subseteq C_i$, A_i can be any subset of C_i, including the full set and the empty set.

B Human Evaluations

We further validate the effectiveness of our proposed methods through human evaluation. The annotator's payment is adequate given the participants' demographic. The annotators are informed of how the data would be used.

B.1 Model-Human Agreement

First we examine the effectiveness of our consistency reward model. We recruit a graduate proficient in English and Chinese, and provide him with a sample of 100 instances from the validation set for each domain, use the model trained with vanilla RL to generate response, and ask him to evaluate the consistency between reasoning trajectory and answer. The Pearson correlation coefficient between human and DeepSeek-V3 is 0.610 with $p = 4.2 \times 10^{-10}$ averaged on two domains, demonstrating LLMs can effectively detect inconsistencies, with a relatively high correlation between model and human.

B.2 Quality Evaluation

Training Data Quality Evaluation We conduct human evaluation on training data to further validate its quality. We randomly sample 50 samples from the augmented dataset obtained from Algorithm 1 for each domain, and examine whether the data quality from two aspects: whether the statement natural, fluent, and easy for humans to understand, and whether the key information in the content between the propositions and original data is the same (i.e. The augmented problem is a well-defined reformulation of the original problem.) We find all 50 samples are well-presented without any incoherence, and 49 in 50 samples maintains all the key information compared with the original data. See Appendix C.2 for training dataset examples. All training data are used in accordance with their original intent and licenses.

Response Quality Evaluation We provide the details of evaluation metrics below. See evaluation results and discussion in Section 5.5.

Correctness (5 points): Focuses on whether the reasoning and conclusions are logically sound and internally consistent. The emphasis is on logical validity rather than domain-specific factual precision.

Professionalism (5 points): Evaluates the accuracy and appropriateness of the domain knowledge used. The key concern is whether the cited knowledge is correct and meaningfully contributes to the reasoning process.

Readability (5 points): Assesses how clearly and smoothly the response can be read and understood, reflecting overall coherence and linguistic fluency.

C More Implementation Details

C.1 Training Configuration

When using our proposed 2-stage training pipeline with consistency reward, for all experiments we uniformly samples 500 instances from the training dataset for stage 1, training for 2 epochs with only answer reward and format reward, and use the rest instances for stage-2 training, with 8 epochs using our consistency reward, answer reward and format reward. Additionally, we change the training data into our shuffled and diversified dataset in stage-2 training after 100 steps when using our data augmentation approach.

The training was conducted on a single node equipped with 8 A800 GPUs using the veRL library (Sheng et al., 2025). The key training configuration is shown below in Table 6. The detailed reward weight is listed in Table 7. Note that we do not conduct hyperparameter searching, so the current performance is underoptimal, and has the potential of further imperovement.

C.2 Prompt and Dataset Example

Data Augmentation We provide the prompts used in augmenting the original dataset, and the instance example of the original dataset and the augmented dataset. The original dataset example of Jec-QA is shown in Table 12, and MedQA in Table 13. Then we concatenate the original statement with options and conduct polish and diversify using LLM. Table 9 shows the prompt used for

Category	Parameter	Value
General	Advantage estimator	GRPO
	Gamma (γ)	1
	Lambda (λ)	1
	Batch size	128
	Max prompt length	1024
	Gradient checkpointing	Enabled
Actor	Learning rate	1×10^{-6}
	Mini-batch size	1024
	Dynamic batch size	Enabled
	KL penalty role	Loss
	KL loss type	Low-variance KL
	KL loss coefficient (β)	0.001
	Entropy coefficient	0.001
	Clip ratio	0.2
	Gradient clipping	1.0
	Sequence parallel size	Model-specific
Rollout	Backend	vLLM
	Tensor model parallel size	2
	Rollouts per sample	6
	Nucleus sampling p	1.0
	GPU memory utilization	0.4
	Sampling temperature	1.0
Qwen2.5-7B	Max response length	2048

Table 6: Training configurations. We use VeRL 0.4.1.dev version.

Phase	Value$_{True}$	Value$_{False}$	Weight
Answer Reward			
Vanilla RL	1.0	0.0	1.0
stage 2	1.0	0.0	1.0
Format Reward			
Vanilla RL	1.0	0.0	1.0
stage 1	1.0	0.0	1.0
stage 2	0.0	-1.0	0.1
Structure Reward			
Stage-1	1.0	0.0	1.0
Consistency Reward			
Stage-2	0.0	-1.0	1.0

Table 7: Reward weights. The final reward score equals to Weight*Value$_{True/False}$

		%Pass Rate	Acc^{+}	Cons	Acc
Standard RL		_	51.4	75.3	61.1
+ *Random*	Data$_{\text{LLM}}$	22.0	49.7	80.0	60.1
+ *Pass-Rate*	Data$_{\text{StrCat}}$	13.8	53.6	78.8	59.0
	Data$_{\text{LLM}}$	14.5	**54.9**	**82.2**	**61.1**

Table 8: Ablation study on data reformulation pipeline. *%Pass Rate* denotes the initial pass rate on Qwen2.5-7B-Instruct.

data reformulation. The final data example after our pipeline is shown in Table 14 and Table 15.

2-stage Consistency Training and Evaluation During 2-stage consistency training, we use a special reward model to monitor the CoT consistency. We use 3-shot in-context learning to instruct the LLM to parse the believed-correct options in the CoT using the Vllm library (Kwon et al., 2023). The prompt for the consistency reward model is shown in Table 10. We also provide the LLM-as-a-judge prompt used in evaluating the quality of the generated CoTs in Table 11.

D Additional Experimental Results

D.1 Ablation Study on Data Reformulation

As shown in Table 1 and Table 2, the proposed dynamic data reformulation approach significantly improves model accuracy across various formats and content. We conduct ablations to assess the impact of: (1) LLM-based diversified data refinement (training with only string-level concatenated data), and (2) pass-rate-based reformulation (using original data without reformulation, or using randomly deconstructed data). The results, shown in Table D.1, reveal that removing LLM-based refinements decreases performance, underscoring the importance of using LLMs to eliminate redundant expressions and diversify content. Additionally, reformulating easy data during training yields the best performance, confirming the effectiveness of our method and supporting prior work on introducing more challenging instances in RL training (An et al., 2025; Huang et al., 2025; Li et al., 2025b).

Notably, consistency also improves with reformulation compared to standard RL, suggesting that curating more challenging data mitigates inconsistency, as tasks requiring complex reasoning naturally encourage models to reason more effectively.

D.2 Detailed Training Dynamics

We present the more detailed dynamics in RL training. In Figures 7 and 8, we report the log of rewards during training, including the answer reward, the consistency reward and the format reward.

D.3 Results For Rule-Based Consistency Reward Model

We explore the feasibility of using a rule-based consistency reward model that leverages exact pattern matching of regular expressions. We find using the rule-based consistency reward results in a performance drop of 11.4% in answer accuracy on the

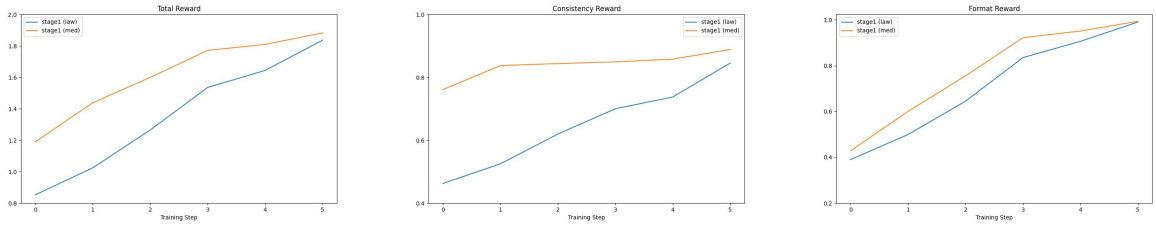

Figure 7: Training dynamics of CLARITY in Stage-1.

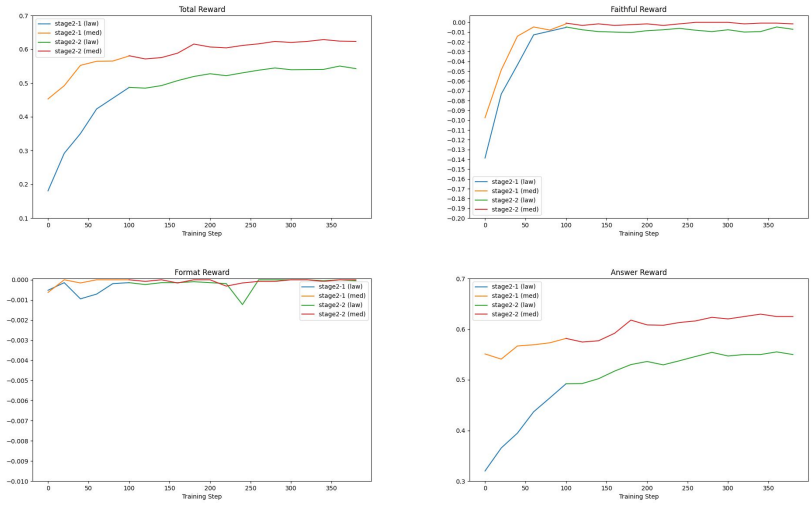

Figure 8: Training dynamics of CLARITY in Stage-2.

validation set, and shows almost no improvement in out-of-domain tasks. The rule-based approach forces the model to generate responses with a fixed format designed to match predefined patterns rather than encouraging deeper reasoning, which makes the model more prone to overfitting during training, producing shorter responses that overly focused on matching the regular expression.

E Case Study

We present case studies from both domains in Table 16 and Table 17. The vanilla Qwen2.5-7B-Instruct model already demonstrates basic legal and medical reasoning capabilities, allowing it to perform limited analytical reasoning. However, its accuracy remains low, often relying on incorrect or incomplete knowledge. Moreover, its responses contain redundant and shallow analyses, suggesting that the distilled reasoning patterns it learned are largely superficial—capturing only surface structures rather than genuine problem-understanding or reasoning depth.

After reinforcement learning, the model's accuracy improves. However, as illustrated in the tables, the standard outcome-based RL introduces inconsistencies: the model may reject a candidate option during reasoning but later select it as the final answer. This inconsistency highlights that outcome-based rewards alone are insufficient to ensure high-quality reasoning.

CLARITY address this issue through a consistency-aware mechanism that explicitly monitors reasoning alignment between intermediate judgments and final conclusions. Additionally, its two-stage refine-then-monitor pipeline further enhances response quality, mitigating potential reward hacking that arises when small-scale reward models fail to detect inconsistencies. When the refine-then-monitor stage is removed (*W/o Stage-1* in the tables), reasoning quality significantly declines, as the model tends to favor oversimplified reasoning patterns even for questions requiring substantial analytical depth.

Data Polishing Prompt Template

I am a {legal/medical} school professor preparing final exam questions for my {legal/medical} course. I would like to design a statement analysis question, where students are required to judge whether the given statement is correct and provide supporting reasoning. To do this, I select a question stem from my multiple-choice question bank and combine it with one of its answer options to transform it into a statement for analysis. However, the stem may still contain phrasing specific to multiple-choice format (e.g., "Which of the following is . . . ?") or redundant punctuation, and may result in illogical reasoning. Therefore, I need your help to polish the statement: remove any phrasing related to multiple-choice format, adjust sentence structure or add connecting words if necessary, and ensure that the final version is logically coherent and easy to understand. You should also diversify the statement by adjusting word order, replacing synonyms or conjunctions, or introducing fictional names and places that do not alter the meaning. This will help diversify the question bank. Please enclose the polished statement within curly braces.

`{Examples}`

Original statement: `{Original_statement}`

Table 9: Prompt for dynamic data reformulation. The original prompt for Jec-QA is written in Chinese, because Jec-QA is presented in Chinese.

Consistency Parsing Prompt Template

You are a {legal/medical} expert. You will be given a multiple-choice {legal/medical}-exam question and a student's line of reasoning. Based on the student's reasoning, identify the final answer selected by the student. Follow the format in the examples below and enclose the answer the student believes is correct in curly braces, e.g. "", "B" or "ABD".

`{Examples}`

Now, based on the examples above, analyze the following question and the student's reasoning, and state the answer the student believes is correct:
Question: `{Question}`
Reasoning: `{Chain-of-Thought}`

Table 10: Prompt for parsing the answer for consistency reward model. The original prompt for Jec-QA is written in Chinese, because Jec-QA is presented in Chinese.

LLM-as-a-judge Prompt Template

You are a medical expert. Below, you will find a medical exam question along with two responses provided by two candidates. Please compare the quality of these two responses based on various factors, including but not limited to: the rationality of the answer, readability, logical coherence, clarity, and ease of understanding.
Please evaluate both answers and determine which one is of higher quality. Use your expertise to carefully analyze the responses. Your final answer should highlight the better response by wrapping it in curly braces, either ¨Answer 1¨ or ¨Answer 2.¨

Question: `{Question}`
Answer 1: `{answer_1}`
Answer 2: `{answer_2}`

Table 11: Prompt for LLM-as-a-judge.

问题: 甲公司与乙公司就双方签订的加工承揽合同达成仲裁协议，约定一旦合同履行发生纠纷，由当地仲裁委员会仲裁。后合同履行发生争议，甲公司将乙公司告上法庭。对此乙公司没有向受诉法院提出异议。开庭审理中，甲公司举出充分证据，乙公司败诉几成定局，于是乙公司向法院提交了双方达成的仲裁协议。法院审查后认为该仲裁协议无效，此时应如何处理? `Question:` Company A and Company B signed a processing contract and reached an arbitration agreement, stipulating that any disputes arising from contract performance would be arbitrated by the local arbitration commission. Later, a dispute arose during contract performance, and Company A sued Company B in court. Company B did not raise any objections to the court's jurisdiction. During the trial, as Company A presented sufficient evidence, making Company B almost certain to lose, Company B submitted the arbitration agreement to the court. After reviewing it, the court found the arbitration agreement invalid. How should the case be handled at this point?	
`Options:` **A**: 继续审理 **B**: 判决该仲裁协议无效 **C**: 如甲公司对仲裁协议效力没有异议，则裁定驳回起诉 **D**: 将仲裁协议的效力问题移交有关仲裁委员会审理	**A**: Continue the trial. **B**: Declare the arbitration agreement invalid. **C**: If Company A does not object to the validity of the arbitration agreement, dismiss the lawsuit by ruling. **D**: Transfer the issue of the arbitration agreement's validity to the relevant arbitration commission for review.

Table 12: An example of Jec-QA case analysis questions. Solving such question requires the model to clarify legal relationships, identify applicable laws, and conduct comprehensive reasoning in complex scenarios rather than rote memorization.

`Question:` A 23-year-old man comes to the physician for evaluation of decreased hearing, dizziness, and ringing in his right ear for the past 6 months. Physical examination shows multiple soft, yellow plaques and papules on his arms, chest, and back. There is sensorineural hearing loss and weakness of facial muscles bilaterally. His gait is unsteady. An MRI of the brain shows a 3-cm mass near the right internal auditory meatus and a 2-cm mass at the left cerebellopontine angle. The abnormal cells in these masses are most likely derived from which of the following embryological structures?
`Options:` **A**: Neural tube **B**: Surface ectoderm **C**: Neural crest **D**: Notochord **E**: Mesoderm

Table 13: An example of MedQA-USMLE questions. These questions assess the model's ability to apply knowledge, concepts, and principles, and the ability to demonstrate fundamental patient-centered skills.

问题**:** 关于民事诉讼法、行政法与行政诉讼法等法律法规的相关内容，下列选项中错误的说法是? **Question:** On matters related to the Civil Procedure Law, the Administrative Law, the Administrative Litigation Law, and other relevant laws and regulations, which of the following statements is incorrect?	
选项： **A**: 5月18日，某市第三中学发生学生集体食物中毒事件，根据《突发公共卫生事件应急条例》的相关规定，第三中学在事发后2小时内向市卫生局报告的处理措施符合法律要求。 **B**: 某商场甲与某电视机生产厂家乙因货款问题产生纠纷，乙遂拒绝向甲商场供货。当客户向该商场销售员询问是否有乙厂的电视机时，销售员故意宣称："乙厂的产品不合格，返修率高，为保护消费者利益，我商场已拒绝销售乙厂的产品。"根据《反不正当竞争法》相关规定，甲商场销售员的上述行为属于诋毁乙厂商业信誉的不正当竞争行为。 **C**: 赵某向陈某借一手机使用，后来陈某向赵某讨要时，赵某表示早就已经将手机还给陈某了，两人因此发生了纠纷。陈某诉至镇上法庭，法庭采用简易程序审理了本案。根据相关法律规定，法庭当庭宣判，告知当事人自己来领取裁判文书，法院将不去送达的做法是正确的。 **D**: 阳光科技公司与星辰制造公司就双方签订的设备加工合同达成仲裁协议，约定若在合同履行过程中出现任何纠纷，均由所在地仲裁委员会进行仲裁。后来在合同执行阶段双方产生争议，阳光科技公司直接将星辰制造公司诉至法院，而星辰制造公司未在法定期间内向受理法院提出管辖权异议。在案件审理过程中，阳光科技公司提供了完整的证据链，星辰制造公司面临败诉风险，此时才向法院提交了双方之前签订的仲裁协议。经法院审查认定该仲裁协议不具备法律效力，在此情况下，法院应当继续审理本案。	Options: **A**. On May 18, a mass food-poisoning incident occurred at a city's No. 3 Middle School. Under the Regulations on Emergency Response to Public Health Emergencies, the school's action of reporting to the municipal health bureau within two hours after the incident complies with legal requirements. **B**. A department store (Party A) and a TV manufacturer (Party B) had a payment dispute, and Party B then refused to supply goods to Party A. When a customer asked the store's salesperson whether the store carried Party B's televisions, the salesperson deliberately stated: "Party B's products are substandard and have a high repair rate. To protect consumers' interests, our store has refused to sell Party B's products." Under the Anti-Unfair Competition Law, the salesperson's conduct constitutes unfair competition by disparaging Party B's commercial reputation. **C**. Zhao borrowed a mobile phone from Chen. Later, when Chen asked for it back, Zhao claimed he had already returned it, leading to a dispute. Chen sued in the town-level people's court, which tried the case under the summary procedure. The court pronounced the judgment in court and told the parties to come and collect the written judgment themselves, stating that the court would not effect service. According to the relevant laws, this practice is correct. **D**. Sunshine Technology Co. and Starlight Manufacturing Co. entered into an arbitration agreement regarding their equipment-processing contract, stipulating that any disputes arising during performance of the contract shall be submitted to the local arbitration commission for arbitration. During performance, a dispute arose and Sunshine Technology directly sued Starlight Manufacturing in court. Starlight did not raise a jurisdictional objection with the court within the statutory period. In the course of the proceedings, Sunshine Technology produced a complete chain of evidence and Starlight faced the risk of losing; only then did Starlight submit the previously concluded arbitration agreement to the court. Upon review, the court determined that the arbitration agreement was not legally valid. In this situation, the court should continue hearing the case.

Table 14: Data example of Jec-QA after applying our augmentation.

Question: Which of the following statements are wrong:
Options: **A**: A startup is working on a novel project in which they claim they can replicate the organelle that is defective in MELAS syndrome. If their project is to mimic the metabolic processes of this organelle, they must be able to replicate fatty acid synthesis.   **B**: Researchers are experimenting with hormone levels in mice in fasting and fed states. To test hormone levels in the fed state, the mice are given an oral glucose load and various hormones are measured in a blood sample. Researchers are most interested in the hormone whose blood levels track evenly with C-peptide levels. Based on the experimental setup, the hormone the researchers are most interested in is responsible for fatty acid breakdown in the body.   **C**: James is a 23-year-old man residing in Boston who presents to his physician for evaluation of progressive hearing loss, dizziness, and tinnitus in his right ear over the past six months. During physical examination, multiple soft, yellowish plaques and papules are noted on his arms, chest, and back. Audiometric testing reveals sensorineural hearing loss, and neurological examination demonstrates bilateral facial muscle weakness along with an unsteady gait. Brain MRI reveals two masses: a 3-cm lesion near the right internal auditory meatus and a 2-cm mass at the left cerebellopontine angle. Given these clinical and imaging findings, the abnormal cells comprising these masses are most likely derived from the neural tube.   **D**: Emma is a 45-year-old woman with a known history of hypertension and bipolar disorder. She visits your clinic today due to new-onset tremors, along with complaints of intense thirst and frequent urination. While her bipolar disorder had been well-controlled with her previous medication regimen, she recently introduced a new drug. Given the symptoms she is experiencing, the medication she most likely started is valproate.   **E**: A 27-year-old man is running on the treadmill at his gym. His blood pressure prior to beginning his workout was 110/72. Based on the physiological response to exercise, a decreased heart rate may be seen in this man now that he is exercising.

Table 15: Data example of MedQA after applying our augmentation.

Algorithm 1 CLARITY

Input: Dataset $\mathcal{D} = \{(Q_i, C_i, Ans_i)\}_{i=0}^{N-1}$, where $C_i = \{c_{i1}, \ldots, c_{iM}\}$, $Ans_i \subseteq C_i$, shuffle proportion $\alpha \leq 1$, statement template set S, Reward weight α_1, β_1, α_2, β_2, γ, shuffle proportion α_{shuf}, number of iterations in both stages

// Dynamic Data Reformulation

for each instance $I_i = (Q_i, C_i, Ans_i)$ in $\mathcal{D}$ **do**
  for each option c_{ij} in C_i **do**
    $c_{ij}^{Div} = \text{LLM}_{Refine}(\text{StrCat}(Q_i, c_{ij}))$
    $Correct_{c_{ij}^{Div}} = \mathbb{I}_{\,c_{ij} \in Ans_i}$
  end for
end for
Descending sort dataset $\mathcal{D}$ by Qwen2.5-7B-Instruct initial pass rate
$\mathcal{D}^{\text{Shuffling}} = \mathcal{D}[0 : \alpha N]$
$C^{\text{Shuffling}} = \bigcup_{i=0}^{\alpha N - 1} C_i$

// Random Grouping

while $C^{\text{Shuffling}}$ is not empty **do**
  $C^{'}$ = Randomly sample M candidates from $C^{\text{Shuffling}}$ without replacement
  $Ans^{'} = \{C \mid C \in C^{'}, Correct(C) = 1\}$
  $Q^{'}$ = Randomly sample a template from S
  $I^{'} = (Q^{'}, C^{'}, Ans^{'})$
  $\mathcal{D}^{'}$.append($I^{'}$)
end while
$\mathcal{D}_{\text{new}} = \text{RandomShuffle}(\mathcal{D}^{'} + \mathcal{D}[\alpha N :])$

// Training Stage-1: Refining

for iteration in Stage 1 **do**
  Generate responses o_i
  for each response o_i **do**
    $\text{Thinking}_{\text{gen}}, \text{Ans}_{\text{gen}} = \text{Split}(o_i)$
    $R_{\text{structure}} = \bigcap_{i=1}^{M} \mathbb{I}_{\text{re.find}(\text{chr}(A+i), \text{Thinking}_{\text{gen}})}$
    $R = \alpha_1 R_{\text{format}} + \beta_1 R_{\text{structure}}$
    Compute advantage $\hat{A}_i$ and optimize $\mathcal{J}_{GRPO}(\theta)$
  end for
end for

// Training Stage-2: Monitoring

for iteration in Stage 2 **do**
  Generate responses o_i
  for each response o_i **do**
    $\text{Thinking}_{\text{gen}}, \text{Ans}_{\text{gen}} = \text{Split}(o_i)$
    $\text{Ans}_{\text{parsed}} = \text{LLM}_{\text{parse}}(\text{Thinking}_{\text{gen}})$
    $R_{\text{consistency}} = \mathbb{I}_{Ans_{\text{parsed}} = Ans_{\text{gen}}}$
    $R_{\text{ans}} = \mathbb{I}_{Ans_{\text{gen}} = Ans_{\text{ref}}}$
    $R = \gamma R_{\text{format}} + \alpha_2 R_{\text{consistency}} + \beta_2 R_{\text{ans}}$
    Compute advantage $\hat{A}_i$ and optimize $\mathcal{J}_{GRPO}(\theta)$
  end for
end for

问题**:** 甲公司借用乙公司的一套设备，在使用过程中不慎损坏一关键部件，于是甲公司提出买下该套设备，乙公司同意出售。双方还口头约定在甲公司支付价款前，乙公司保留该套设备的所有权。不料在支付价款前，甲公司生产车间失火，造成包括该套设备在内的车间所有财物被烧毁。对此，下列哪些选项是正确的? Question: On matters related to the Civil Procedure Law, the Administrative Law, the Administrative Litigation Law, and other relevant laws and regulations, which of the following statements is incorrect?	
选项： **A**: 乙公司已经履行了交付义务，风险责任应由甲公司负担 **B**: 在设备被烧毁时，所有权属于乙公司，风险责任应由乙公司承担 **C**: 设备虽然已经被烧毁，但甲公司仍然需要支付原定价款 **D**: 双方关于该套设备所有权保留的约定应采用书面形式 Options: **A**: Company B has fulfilled its delivery obligation, and the risk responsibility should be borne by Company A. **B**: When the equipment was destroyed, the ownership remained with Company B, and the risk responsibility should be borne by Company B. **C**: Although the equipment has been destroyed, Company A is still required to pay the original price. **D**: The agreement between both parties regarding the retention of ownership of the equipment should be in writing. Answer: **AC**	
Qwen2.5-7B-Instruct	
<思考> 本题涉及所有权保留买卖合同中的风险责任分配问题。根据《合同法》第142条，标的物毁损、灭失的风险，在标的物交付之前由出卖人承担，交付之后由买受人承担，但法律另有规定的除外。本题中，甲公司与乙公司口头约定在甲公司支付价款前，乙公司保留该套设备的所有权，因此在甲公司支付价款前，设备的所有权仍属于乙公司。根据《合同法》第142条，设备被烧毁的风险应由乙公司承担。选项A错误，选项B正确。根据《合同法》第167条，当事人约定标的物所有权保留的，出卖人对标的物保留的所有权，未经登记，不得对抗善意第三人。本题中，甲公司与乙公司未采用书面形式，因此该约定不能对抗善意第三人。选项D正确。根据《合同法》第142条，设备被烧毁的风险应由乙公司承担，因此甲公司无需支付原定价款。选项C错误。</思考><回答> 本题答案为B和D。 </回答>	<Thought> This question concerns risk allocation in a retention of title sales contract. According to Article 142 of the Contract Law, the risk of damage or loss of the subject matter is borne by the seller before delivery and by the buyer after delivery, unless otherwise specified by law. Here, Company A and Company B orally agreed that Company B retains ownership of the equipment until Company A pays, so ownership remains with Company B until payment. According to Article 142, the risk of destruction is borne by Company B. Option A is incorrect, and option B is correct. According to Article 167 of the Contract Law, a retention of title agreement must be registered to be enforceable against bona fide third parties. Since the agreement was not in writing, it cannot be asserted against a bona fide third party. Therefore, option D is correct. As the risk of destruction is borne by Company B, Company A does not need to pay the original price. Option C is incorrect.</Thought> <Answer> The correct answers are B and D. </Answer>
Vanilla Reinforcement Learning	
<思考> 根据《中华人民共和国合同法》的规定，标的物毁损、灭失的风险，在标的物交付之前由出卖人承担，交付之后由买受人承担，但法律另有规定的除外。在本案例中，甲公司和乙公司之间存在设备买卖合同，且双方口头约定在甲公司支付价款前，乙公司保留设备的所有权。因此，设备在交付给甲公司后，风险责任应由甲公司承担。但是，由于设备被烧毁，设备的所有权归属问题需要明确。根据《中华人民共和国合同法》第一百三十四条的规定，当事人可以在买卖合同中约定买受人未履行支付价款或者其他义务的，标的物的所有权属于出卖人。因此，双方关于设备所有权保留的约定是有效的。但是，由于设备被烧毁，设备的价值已经不存在，因此甲公司无需支付原定价款。 选项A：根据上述分析，设备在交付给甲公司后，风险责任应由甲公司承担，因此该选项正确。 选项B：由于设备在交付给甲公司后，风险责任应由甲公司承担，因此该选项错误。 选项C：由于设备被烧毁，设备的价值已经不存在，因此甲公司无需支付原定价款，因此该选项错误。 选项D：根据《中华人民共和国合同法》第一百三十四条的规定，当事人可以在买卖合同中约定买受人未履行支付价款或者其他义务的，标的物的所有权属于出卖人。因此，双方关于设备所有权保留的约定应采用书面形式，因此该选项正确。综上所述，正确的选项是A、C。</思考><回答> AC </回答>	<Thought> According to the Contract Law of the People's Republic of China, the risk of damage or loss of the subject matter is borne by the seller before delivery and by the buyer after delivery, unless otherwise specified by law. In this case, there is a sales contract between Company A and Company B, with an oral agreement that Company B retains ownership of the equipment until Company A pays. Therefore, after delivery, the risk is borne by Company A. However, as the equipment was destroyed, the issue of ownership needs to be clarified. According to Article 134 of the Contract Law, the parties may agree that ownership remains with the seller if the buyer fails to pay the price or fulfill other obligations. Therefore, the retention of ownership agreement is valid. However, since the equipment is destroyed, its value no longer exists, and Company A does not need to pay the original price. Option A: Based on the above analysis, after delivery, the risk responsibility is borne by Company A, so this option is correct. Option B: Since the risk responsibility is borne by Company A after delivery, this option is incorrect. Option C: Since the equipment is destroyed and its value no longer exists, Company A does not need to pay the original price, so this option is incorrect. Option D: According to Article 134 of the Contract Law, the retention of ownership agreement must be in writing, so this option is correct. Therefore, the correct answers are A and C. </Thought> <Answer> AC </Answer>
W/o Stage-1	
<思考> 该题涉及所有权保留买卖合同中的风险责任和价款支付问题。正确答案是A和C。A选项正确，因为设备已经交付，风险责任由买方承担。C选项正确，即使设备毁损，甲公司仍需支付原定价款。B选项错误，因为设备毁损时所有权属于乙公司，风险责任应由甲公司承担。D选项错误，关于所有权保留的约定可以采用口头形式。因此，答案是AC。 </思考><回答> AC </回答>	<Thought> This question involves risk responsibility and payment issues in a retention of title sales contract. The correct answers are A and C. Option A is correct because the equipment has been delivered, and the risk is borne by the buyer. Option C is correct because Company A must still pay the original price, even if the equipment is damaged. Option B is incorrect because ownership was with Company B when the equipment was damaged, and the risk responsibility should be borne by Company A. Option D is incorrect because the retention of title agreement can be oral. Therefore, the answer is AC. </Thought> <Answer> AC </Answer>
CLARity	
<思考> 本题考查所有权保留买卖中的风险承担问题。 根据《合同法》第一百四十一条规定，出卖人交付标的物后，风险责任一般由买受人承担。但是，本题中甲公司和乙公司有特别约定，在甲公司支付价款前，乙公司保留该套设备的所有权。因此，根据《合同法》第一百四十二条规定，标的物毁损、灭失的风险，在标的物交付之后由买受人承担，但法律另有规定或者当事人另有约定的除外。本题中，风险责任应由甲公司承担，因此选项A正确，选项B错误。 根据《合同法》第一百四十七条规定，出卖人交付标的物后，买受人在标的物毁损、灭失的风险发生后尚未支付价款的，出卖人有权要求买受人支付全部价款或者就标的物的毁损、灭失获得赔偿。因此，即使设备被烧毁，甲公司仍然需要支付原定价款，因此选项C正确。 根据《合同法》第一百六十七条规定，出卖人保留所有权的，应当与买受人约定所有权保留的期间。因此，关于所有权保留的约定应当采用书面形式，因此选项D错误。 因此，正确的选项是A和C。AC </思考><回答> AC </回答>	<Thought> This question examines the issue of risk allocation in a retention of title sales contract. According to Article 141 of the Contract Law, after the seller delivers the subject matter, the risk is generally borne by the buyer. However, in this case, Company A and Company B have a special agreement that Company B retains ownership of the equipment until Company A pays. Therefore, according to Article 142 of the Contract Law, the risk of damage or loss is borne by the buyer after delivery, unless otherwise specified by law or agreed by the parties. In this case, the risk is borne by Company A, so Option A is correct, and Option B is incorrect. According to Article 147 of the Contract Law, if the buyer has not paid the price after the risk of damage or loss occurs, the seller has the right to demand full payment or compensation for the damage or loss. Therefore, even if the equipment is destroyed, Company A must still pay the original price, so Option C is correct. According to Article 167 of the Contract Law, if the seller retains ownership, the period of retention must be agreed upon with the buyer. Therefore, the retention of title agreement should be in writing, making Option D incorrect. Thus, the correct answers are A and C. AC </Thought> <Answer> AC </Answer>

Table 16: A model response example from different methods in Jec-QA validation set. Red part contains inconsistency.

`Question:` Question: A 33-year-old man presents to the emergency department with joint pain. He states that since yesterday he has had joint pain, and today he noticed a rash. The patient does not have a significant past medical history. The patient smokes cigars, drinks 1 to 4 drinks per day, is sexually active, and uses cocaine occasionally. His temperature is 101°F (38.3°C), blood pressure is 125/65 mmHg, pulse is 80/min, respirations are 14/min, and oxygen saturation is 98% on room air. Physical exam reveals pustular skin lesions and an inflamed and painful right knee that is warm to the touch. Which of the following is the most appropriate treatment for the most likely diagnosis? `Answer:` **B**
`Options:` **A**: Azithromycin **B**: Cefixime and azithromycin **C**: Ceftriaxone **D**: Ceftriaxone and vancomycin **E**: Piperacillin and tazobactam
Qwen2.5-7B-Instruct
<think> To determine the most appropriate treatment, we need to consider the patient's symptoms and the likely diagnosis. The patient presents with joint pain and a rash, which could indicate a variety of conditions, but the presence of pustular skin lesions and an inflamed, warm, and painful knee suggests a possible bacterial infection, such as gonorrhea or chlamydia, which can present with these symptoms. - **Azithromycin (A)**: This is an antibiotic commonly used for chlamydia and other bacterial infections. However, it may not be sufficient for a more severe infection like gonorrhea. - **Cefixime and azithromycin (B)**: Cefixime is a cephalosporin antibiotic that is effective against a wide range of bacteria, including those causing gonorrhea. Azithromycin is also included to cover for chlamydia. This combination is often used for treating sexually transmitted infections (STIs). - **Ceftriaxone (C)**: This is a broad-spectrum cephalosporin that is highly effective against gonorrhea. It is often used as a single-dose treatment for gonorrhea. - **Ceftriaxone and vancomycin (D)**: Ceftriaxone is effective against gonorrhea, but adding vancomycin is typically not necessary unless there is a concern for resistant strains or other complications. - **Piperacillin and tazobactam (E)**: This combination is used for more severe infections, such as those caused by gram-negative bacteria, but it is not the first-line treatment for the likely diagnosis in this case. Given the symptoms and the likely diagnosis of gonorrhea, the most appropriate treatment would be a combination that effectively covers both the likely bacterial cause and any potential co-infections. </think><answer> The most likely diagnosis is gonorrhea, given the symptoms and the patient's sexual activity. The most appropriate treatment would be a combination of Ceftriaxone and Azithromycin (B) to cover both the bacterial infection and the potential for co-infection with chlamydia. </answer>B
Vanilla Reinforcement Learning
<think> To determine the most likely diagnosis and appropriate treatment, let's analyze the symptoms and physical exam findings. The patient has joint pain, a rash, and pustular skin lesions, which are consistent with gonococcal arthritis and dermatitis (gonococcal infection). Gonococcal infection is caused by Neisseria gonorrhoeae and can present with these symptoms. The most likely diagnosis is gonococcal arthritis and dermatitis. The treatment for gonococcal infection is usually ceftriaxone, which is a broad-spectrum antibiotic effective against Neisseria gonorrhoeae. Let's consider the treatment options: A: Azithromycin - This is an antibiotic used for chlamydia and other infections, but not for gonorrhea. B: Cefixime and azithromycin - Cefixime is an oral cephalosporin, but it is not as effective as ceftriaxone for gonorrhea. C: Ceftriaxone - This is the standard treatment for gonorrhea. D: Ceftriaxone and vancomycin - Vancomycin is used for methicillin-resistant Staphylococcus aureus (MRSA) and other Gram-positive infections, but it is not necessary for gonorrhea. E: Piperacillin and tazobactam - This is a broad-spectrum antibiotic combination used for Gram-negative and Gram-positive infections, but it is not the standard treatment for gonorrhea. </think><answer>D</answer>
W/o Stage-1
<think> To determine the most appropriate treatment, we need to first identify the most likely diagnosis based on the patient's symptoms. The patient presents with pustular skin lesions and a painful, warm knee, which are consistent with gonococcal arthritis or skin infection caused by Neisseria gonorrhoeae or Chlamydia trachomatis. However, the presence of pustular lesions is more indicative of a bacterial skin infection, likely caused by Staphylococcus aureus, particularly methicillin-resistant Staphylococcus aureus (MRSA). Given the patient's history of cocaine use and occasional sexual activity, we need to consider the possibility of sexually transmitted infections (STIs) such as gonorrhea or chlamydia. However, the pustular lesions are more suggestive of a skin infection rather than a STI. The most likely diagnosis is a skin infection caused by MRSA, which is common in individuals with risk factors for skin infections, such as cocaine use and smoking. The treatment for MRSA skin infections typically includes antibiotics effective against both methicillin-susceptible and methicillin-resistant Staphylococcus aureus. The most appropriate antibiotic for MRSA is usually a beta-lactam antibiotic with beta-lactamase inhibition, such as cloxacillin, or a glycopeptide like vancomycin. However, for outpatient treatment, oral antibiotics like clindamycin or doxycycline are often used. Given the options, the most appropriate treatment for a likely MRSA skin infection is: A: Azithromycin - This is typically used for chlamydia and mycoplasma infections, not for MRSA. B: Cefixime and azithromycin - Cefixime is not effective against MRSA, and azithromycin is not the first-line treatment for MRSA. C: Ceftriaxone - This is typically used for gonorrhea and chlamydia, not for MRSA. D: Ceftriaxone and vancomycin - Ceftriaxone is not effective against MRSA, and vancomycin is used for severe cases or when MRSA is suspected. E: Piperacillin and tazobactam - This is typically used for Gram-negative infections and is not the first-line treatment for MRSA. The most appropriate treatment for a likely MRSA skin infection is a beta-lactam antibiotic with beta-lactamase inhibition, such as dicloxacillin or a glycopeptide like vancomycin. However, among the options provided, the most appropriate treatment would be a combination of an antibiotic effective against MRSA and a broad-spectrum antibiotic to cover other potential infections. Given the options, the most appropriate treatment is Ceftriaxone and vancomycin, which is typically used for severe cases or when MRSA is suspected. Therefore, the most appropriate treatment for the most likely diagnosis is D: Ceftriaxone and vancomycin.</think><answer>D</answer>
CLARity
<think>To determine the most appropriate treatment, we first need to identify the most likely diagnosis based on the patient's symptoms and physical exam findings. The patient presents with fever, joint pain, and pustular skin lesions. These symptoms are characteristic of gonococcal arthritis, which is caused by Neisseria gonorrhoeae. Gonococcal arthritis typically affects large joints and is associated with other signs of gonococcal infection such as skin lesions. The most common approach to treating gonococcal arthritis is to treat the underlying gonorrhea infection. The standard treatment for uncomplicated gonorrhea includes a single intramuscular dose of ceftriaxone. Given that the patient also has skin lesions, which could indicate disseminated gonococcal infection (DGI), a more aggressive treatment regimen is recommended. The Centers for Disease Control and Prevention (CDC) recommends ceftriaxone 250 mg IM plus either azithromycin 1 g orally or cefixime 400 mg orally for the treatment of DGI. Therefore, the most appropriate treatment for this patient would be ceftriaxone and either azithromycin or cefixime. Let's evaluate each option: A: Azithromycin - This is a suitable oral antibiotic for co-treatment with ceftriaxone in DGI. B: Cefixime and azithromycin - This is the recommended combination according to CDC guidelines. C: Ceftriaxone - This is part of the treatment but needs to be combined with an oral antibiotic. D: Ceftriaxone and vancomycin - Vancomycin is not typically used for gonorrhea unless there is evidence of penicillin allergy or resistance. E: Piperacillin and tazobactam - This combination is used for more severe infections like endocarditis but is not the standard treatment for gonococcal arthritis. Given the guidelines, the most appropriate treatment is Ceftriaxone 250 mg IM plus Azithromycin 1 g orally. Therefore, the correct answer is B.</think><answer>B</answer>

Table 17: A model response example from different methods in MedQA validation set. Red part contains inconsistency.