A. Construction of Cold-start Data

As described in Section **??**, we construct the cold-start data through the following procedure:

1. We first apply the policy model to generate answers for all questions and categorize the resulting samples into initially correct and incorrect groups.
2. To ensure high-quality revisions, we additionally query GPT to produce its own answers for the same questions prior to any modification and check whether it can correctly answer the question.
3. For samples where the policy model produces incorrect answers, we use the prompt *Prompt: Revise Incorrect Responses"* to revise the initial responses. This prompt explicitly injects the reasoning elements of clarifying the question, verifying the reasoning, and rethinking.
4. For samples where the policy model's answers are correct, we first evaluate whether an alternative solution exists using GPT-4o with the *Prompt: Alternative Solution Detection"*. If a conceptually different solution is feasible, we employ the prompt *"Prompt: Revise Correct Responses"* to refine the response by incorporating clarification of the question, verification steps, and exploration of alternative approaches.

Prompt: Revise Incorrect Responses

You will be given an image, a question, a reference answer, and a response from another model to revise. Your task is to:

Analysis:

1. Check the result of the given response and state correctness.
2. Identify all reasoning mistakes in the response.

Then modify the response so that it incorporates the following content:

1. Modify the beginning of the response so that it starts from clarifying the question.
   Clarify the Question: restate the goal clearly at the beginning, such as the units of answer required, output format, etc.
2. Keep the rest of the given response with its original wrong result unchanged. The original result should be put in `\boxed{}`.
3. After the original result, add content to verify the above reasoning process.
   Verify – analyze whether there are errors in the reasoning process and verify the original results.
4. If reasoning errors are found, rethink the question.
   Rethink – reasoning step by step again and give a new result.
5. After all reasoning is completed, put the final result in `\boxed{}`.
6. When adding content for "Clarify the Question", "Identify Known Information", "Verify", or "Rethink", adopt the following anchor sentences or similar expressions.

Anchor Sentences or Fragments:

Clarify the Question:

- "To solve the problem, we need to determine the . . . "
- "The question asks us to . . . "
- "To solve this problem, let's restate what is being asked."
- "The task is to identify the . . . "
- "We begin by clarifying the goal of the problem."

Verify:

- "Let's do a quick check to confirm . . . "
- "We can verify this by . . . "
- "We should test the answer against the problem requirements."
- "Let's double-check whether the result satisfies all constraints."
- "We can confirm this by re-evaluating the key steps."
- "Let's cross-check the calculation to make sure we didn't miss anything."
- "We should check if the answer remains consistent with . . . "
- "We can validate this result by comparing . . . "
- "Let's verify the reasoning by examining the . . . "
- "We need to verify this by . . . "

- "Let's run a quick consistency check on the answer."
- "Let's inspect whether our result contradicts any given information."
- "Let's confirm this by calculating the expected outcome."
- "Let's test this value to see if it satisfies ..."

Rethink:

- "Let's re-evaluate ..."
- "This suggests we should reconsider the approach."
- "Let's start over from where we went wrong ..."
- "Let's rethink the steps that led to this result."
- "It seems there may be a mistake, let's reassess the solution."
- "Maybe our initial interpretation is incorrect, let's revise it."
- "Let's go back and analyze the key step that caused the error."
- "This result doesn't look right, let's rethink the strategy."
- "Let's revisit the earlier deduction with more caution."
- "Let's backtrack a bit and look for a more accurate reasoning path."
- "If this appears inconsistent, we should rethink the entire approach."
- "Let's reconsider the conclusion now that we see the error."

Important Guidelines:

- The final revised response should remain coherent with the same language tone. Do not mention the original response, reference answer, or any meta information.
- Do not revise the reasoning content in the original response or the original result. Only the new Verify and Rethink parts should contain additional reasoning.
- Present the final result at the end, formatted as `\boxed{...}` with nothing after it.
- For multiple-choice questions, both the original and final results should be the option letter directly (e.g., A, B, C, D).

Output Format:

- **Analysis:** [1. Correctness of the original answer. 2. Identified reasoning errors. 3. Plan for modification.]
- **Revised Response:**
  - Reasoning the question: [Original Response.]
  - Verify: [Analyze reasoning and verify original results.]
  - Rethink: [Step-by-step reasoning and new result.]

Question: {question}
Reference Answer: {gpt_answer}
Model's Response: {response}

Prompt: Alternative Solution Detection

Given an image, a question, and an answer from another model, analyze the question and the provided answer to determine whether you can solve it using a different approach. If the question is too simple to allow for a conceptually different solution, respond with `No`.

Output Format:

Is a different solution feasible?
[Yes or No]
A different way to solve the question:
[If feasible, provide the complete reasoning process and the final answer in `\boxed{}`.]

Prompt: Revise Correct Responses

You will be given an image, a question, a response from another model to revise, and a reference different solution for this question. Your task is to:

Analysis:

1. Check the result of the given response and state correctness.

Then modify the response so that it incorporates the following content:

1. Modify the beginning of the response so that it starts from clarifying the question.
   Clarify the Question: restate the goal clearly at the beginning, such as the units of answer required, output format, etc.
2. Keep the rest of the given response with its original result unchanged. The original result should be put in `\boxed{}`.
3. After the original result, add content to verify the reasoning process.
   Verify – analyze whether there are errors in the reasoning process and verify the original results. Avoid merely restating the original reasoning. You should validate by, for example, testing boundary conditions or examining whether the answer matches the image content.
4. If reasoning is correct, try a different approach to enrich the reasoning.
   Try Different Approach – provide a different approach and perform detailed reasoning to cross-validate the results.
5. After all reasoning is completed, put the final result in `\boxed{}`.
6. When adding content for "Clarify the Question", "Verify", or "Try Different Approach", adopt the following anchor sentences or similar expressions.

Anchor Sentences or Fragments:

Clarify the Question:

- "To solve the problem, we need to determine the ..."
- "The question asks us to ..."
- "To solve this problem, let's restate what is being asked."
- "The task is to identify the ..."
- "We begin by clarifying the goal of the problem."

Verify:

- "Let's do a quick check to confirm ..."
- "We can verify this by ..."
- "We should test the answer against the problem requirements."
- "Let's double-check whether the result satisfies all constraints."
- "We can confirm this by re-evaluating the key steps."
- "Let's cross-check the calculation to make sure we didn't miss anything."
- "We should check if the answer remains consistent with ..."
- "We can validate this result by comparing ..."
- "Let's verify the reasoning by examining the ..."
- "We need to verify this by ..."
- "Let's run a quick consistency check on the answer."
- "Let's inspect whether our result contradicts any given information."
- "Let's confirm this by calculating the expected outcome."
- "Let's test this value to see if it satisfies ..."

Try Different Approach:

- "Try another method"
- "Consider an alternative approach"
- "Let's attempt an alternate approach"
- "Apply another way to ..."
- "Explore a secondary route"

Important Guidelines:

- The revised response should remain coherent as a whole with the same language tone. Do not explicitly mention the response, original response, reference solution, correct answer, reference answer, or any other mentions of given content.
- Do not revise the reasoning content in the original response or its original result. Only the newly added Verify and Try-Different-Approach parts should include new reasoning.
- Present the final answer at the end, formatted as `\boxed{...}` with nothing after it.
- For multiple-choice questions, both the original result and final result should be the option letter only (e.g., A, B, C, D, E, F).

Output Format:
Analysis:
[1. State whether the original final answer is correct or incorrect. 2. Propose a plan on how to modify the response.]
Revised Response:
- Reasoning the question: [Original Response.]
- Verify: [Analyze whether there are errors in the reasoning process and verify the original results.]
- Try Different Approach: [Provide a different approach to cross-validate the answer.]

Question: {question}
Reference Answer: {gpt_answer}
Model's Response: {response}

B. Templates for Matching Reasoning Patterns

As shown in Section **??**, we collect a set of template sentences of each reasoning pattern, then leverage the embedding similarity between template sentences and models' responses to match the reasoning patterns. The template sentences are shown as following:

Reasoning Pattern Templates

Clarify the Question:
- To solve the problem, we need to determine the...
- The question asks us to...
- To solve this problem, let's restate what is being asked.
- The task is to identify the...
- We begin by clarifying the goal of the problem.

Verify:
- Let's do a quick check to confirm...
- We can verify this by...
- We should test the answer against the problem requirements.
- To verify...
- Let's verify...
- Let's double-check whether the result satisfies all constraints.
- We can confirm this by re-evaluating the key steps.
- Let's cross-check the calculation to make sure we didn't miss anything.
- We should check if the answer remains consistent with...
- We can validate this result by comparing...
- Let's verify the reasoning by examining the...
- We should confirm whether this aligns with...
- Let's check whether this value fits the...
- We need to verify this by...
- Let's ensure the logic is sound by reviewing the assumptions.
- Let's confirm the calculation...
- Let's run a quick consistency check on the answer.
- To validate this approach, let's re-check the boundary conditions.
- Let's inspect whether our result contradicts any given information.
- Let's confirm this by calculating the expected outcome.
- Let's test this value to see if it satisfies...

Different Approach:
- Try another method.
- Consider an alternative approach.
- Let's attempt an alternate approach.
- Apply another way to...

	MathVista	MathVision	MMMUPro	Average	Response Length
Verify	75.6	34.6	40.5	50.23	834
Clarify + Verify	74.5	34.2	41.2	49.97	958
Clarify + Verify + Explore	75.2	33.6	42.3	50.37	1063
$\alpha = 0.05$	74.1	34.0	42.4	50.17	941
$\alpha = 0.075$	75.0	33.3	42.2	50.16	1041
$\alpha = 0.1$	75.2	33.6	42.3	50.37	1063

Table 1. Ablation Experiments on reasoning patterns and α that controls the weight of reasoning rewards.

- Explore a secondary route.
- This suggests we should reconsider our approach.
- Let's start over from where the logic went wrong.
- Let's rethink the steps that led to this result.
- It seems there may be a mistake, let's reassess the solution.
- Let's reconsider the assumptions we made at the beginning.
- Maybe our initial interpretation is incorrect, let's revise it.
- Let's go back and analyze the key step that caused the error.
- Let's reflect on whether the earlier step aligns with the question.
- This result doesn't look right, let's rethink the strategy.
- Let's revisit the earlier deduction with more caution.
- To avoid the previous mistake, let's restart from the core idea.
- Let's examine whether we misunderstood part of the prompt.
- Let's backtrack a bit and look for a more accurate reasoning path.
- If this appears inconsistent, we should rethink the entire approach.
- Let's reconsider the conclusion now that we see the error.
- Let's re-evaluate the reasoning from earlier.

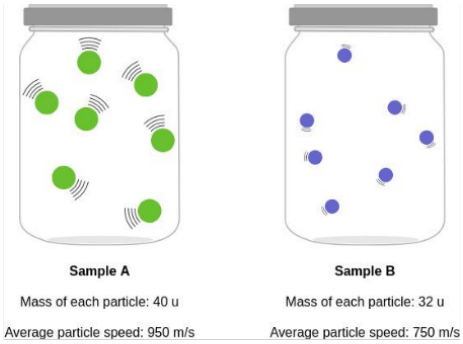
C. Implementation Details

We build our RLRR framework on top of Group Relative Policy Optimization (GRPO). To promote exploration during training, we remove the KL regularization term from the original GRPO objective. We further filter each response group based on group accuracy, discarding queries whose sampled responses are all correct or all incorrect, thereby maintaining the quality and diversity of training signals. When computing the loss, rather than averaging token-level losses over the entire batch, we first compute the average loss for each sequence and then average across the entire batch. This sequence-level normalization ensures that each sample contributes equally during optimization, regardless of its response length.

D. More Experiments

Table 1 analyzes the effect of different reasoning patterns and the weighting factor α on model performance. Using only the verification reward already yields strong results, achieving a 50.23% average score with a relatively short response length. Adding clarification before verification slightly increases output length and improve the accuracy on MMMUPro. Introducing exploration improves performance on MMMU-Pro (from 40.5% to 42.3%), but also leads to the longest responses. These results highlight a trade-off: encouraging more reasoning behaviors can improve robustness on complex benchmarks, but may also induce longer, less concise reasoning traces. We further examine the effect of the reasoning reward weight α. A smaller value ($\alpha = 0.05$) produces stable improvements with moderate output length, while increasing α to 0.075 or 0.1 encourages stronger use of reasoning patterns but results in noticeably longer responses with a slightly better performance. Overall, these findings suggest that verification plays the most critical role in multimodal reasoning, exploration provides complementary benefits for challenging tasks.

The diagrams below show two pure samples of gas in identical closed, rigid containers. Each colored ball represents one gas particle. Both samples have the same number of particles. Compare the average kinetic energies of the particles in each sample. Which sample has the higher temperature?



To determine which sample has the higher temperature, we need to compare the average kinetic energies of the particles in each sample. The average kinetic energy of a gas particle is given by the equation:
\[Kinetic Energy = \frac{1}{2} m v^2 \]
where \(m \) is the mass of the particle and \(v \) is the speed of the particle.
Let's calculate the kinetic energy for each sample:
Sample A:
- Mass of each particle, \(m_A = 40 \, u \)
- Average particle speed, \(v_A = 950 \, m/s \)
\[Kinetic Energy_A = \frac{1}{2} \times 40 \, u \times (950 \, m/s)^2 \]
Sample B:
- Mass of each particle, \(m_B = 32 \, u \)
- Average particle speed, \(v_B = 750 \, m/s \)
\[Kinetic Energy_B = \frac{1}{2} \times 32 \, u \times (750 \, m/s)^2 \]
Now, let's compare these kinetic energies. Since the number of particles is the same in both samples, we can compare the product of the mass and the square of the speed for each sample:
- For Sample A: \(40 \, u \times (950 \, m/s)^2 \)
- For Sample B: \(32 \, u \times (750 \, m/s)^2 \)
Calculating the values:
- \((950 \, m/s)^2 = 902500 \, (m^2/s^2) \)
- \((750 \, m/s)^2 = 562500 \, (m^2/s^2) \)
So,
- Kinetic Energy_A = \(40 \times 902500 = 36100000 \, u \cdot m^2/s^2 \)
- Kinetic Energy_B = \(32 \times 562500 = 18000000 \, u \cdot m^2/s^2 \)
Since Kinetic Energy_A > Kinetic Energy_B, Sample A has the higher average kinetic energy.
The temperature of a gas is directly proportional to the average kinetic energy of its particles. Therefore, Sample A has the higher temperature.
The correct answer is (B).

Let's do a quick check to confirm the reasoning. The calculation of the kinetic energy for each sample is correct, using the formula for kinetic energy and the given values. The comparison of the product of mass and the square of speed confirms that Sample A has the higher kinetic energy. Since temperature is directly proportional to the average kinetic energy, this correctly identifies Sample A as the one with the higher temperature. The reasoning and calculations are consistent with the physics of gas kinetics.

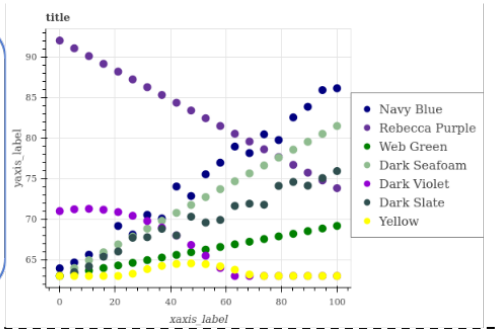
The final answer is \boxed{B}.

Figure 1. Example Response

E. Response Examples

Figure 1, 2, 3 shows some responses from the model. After training, the model is able to selectively apply different reasoning patterns, such as verification or exploring alternative approaches, when solving a problem.

Is Web Green greater than Yellow?



To solve this problem, let's restate what is being asked. The question asks us to determine whether "Web Green" is greater than "Yellow" based on the data presented in the scatter plot. The answer should be a simple yes or no, corresponding to the choices (A) yes or (B) no.

1. **Identify the data points for Web Green and Yellow:**
   - Web Green is represented by the green dots.
   - Yellow is represented by the yellow dots.

2. **Compare the y-values:**
   - The y-values for Web Green start around 65 and increase gradually.
   - The y-values for Yellow start around 63 and remain relatively constant.

3. **Analysis:**
   - For most of the x-values, the y-values of Web Green are higher than those of Yellow.

Given this analysis, we can conclude that Web Green is indeed greater than Yellow for the majority of the data points.

Therefore, the correct answer is:
(A) yes

Let's do a quick check to confirm the reasoning. The scatter plot shows that the Web Green dots (green) are generally positioned above the Yellow dots (yellow) along the y-axis. This visual confirmation supports the conclusion that Web Green is greater than Yellow for the majority of the data points. The y-values for Web Green start at around 65 and increase, while the y-values for Yellow start at around 63 and remain relatively stable. This consistent pattern confirms the correctness of the answer.

Consider an alternative approach to solving the problem. Instead of focusing on the visual inspection, we can also look at the trend lines of the two colors. Web Green appears to have a positive slope, indicating an increasing trend in y-values as x increases. In contrast, Yellow has a nearly horizontal line, indicating a constant y-value. Since the starting y-value for Web Green is higher than that for Yellow, and it increases over time, it is clear that Web Green is greater than Yellow for the majority of the data points. This alternative method confirms the previous conclusion without relying on visual inspection alone.

The final answer is:
\boxed{(B)}

Figure 2. Example Response

what is the total volume of the measuring cup?

To solve this problem, let's restate what is being asked. The question asks us to determine the total volume of the measuring cup in grams. The answer should be an integer value in grams.

To determine the total volume of the measuring cup, we need to look at the markings on the cup. The cup is labeled with measurements in milliliters (ml) and grams (g).

1. The cup is marked with a scale that goes up to 1000 ml.
2. The scale also shows that 1000 ml of water weighs approximately 1000 g.

Therefore, the total volume of the measuring cup is 1000 ml, which is equivalent to 1000 g.

The final answer is:
\boxed{1000}

Figure 3. Example Response