

## A CASE STUDY OF EAI

We conduct a qualitative study how actor and critic accomplish exploration. On GSM8K, we manually inspect 100 examples randomly sampled from EAI explored dataset to better understand how does the exploration work. We observed that the critic excels at assessing the novelty of generated questions, providing precise feedback to enhance them. The actor effectively utilizes the critic’s feedback to refine the generated questions. Of the 100 samples reviewed, in 24 instances, the critic initially found the questions lacking in novelty and provided suggestions for improvement. Out of these, 19 were subsequently refined and accepted after revisions. Below is a representative example of this interaction where the critic offers advice to significantly differentiate the generated question. Here, the critic recommends incorporating the mathematical concept of volumes and presenting an alternative scenario

[Critic guides exploration]

(First round)

Actor’s generation: Question: Jake is planning to buy a new skateboard which costs \$200. Jake only has a quarter of the money he requires. His brother decided to give him \$20 for it, and his uncle three times as much as his brother. How much more does Jake need to raise to purchase the skateboard? (answer omitted)

Critic’s critique: The new question you’ve provided is somewhat similar in structure to the existing questions, as it involves basic arithmetic operations (multiplication, division, addition, and subtraction). While it introduces a different context with hats and balls, the underlying math principles are similar to those used in the existing questions. To make the new question substantially different, you could consider introducing more complex mathematical concepts such as volumes or using a different scenario altogether. ✘

(Second round)

Actor’s generation: An artist is creating a sculpture using various geometric shapes. The sculpture consists of a cube, a sphere, a cylinder, and a pyramid. The cube has a side length of 6 units, the sphere has a radius of 4 units, the cylinder has a height of 8 units and a base radius of 3 units, and the pyramid has a base edge length of 5 units and a height of 7 units. Calculate the total volume of the artist’s sculpture. A: omitted

Critic’s critique: The new question is substantially different from the existing questions. It involves a slightly more complex problem-solving process. It is good to be added to math questions. The decision is accept. ✔

We also conducted a study to determine how effective the critic is at evaluating the correctness of answers. Of the 100 critiques assessed, the critic accurately identified incorrectness in 13 actor generations and offered critiques to correct 10 of them. However, there were 7 instances where the critic failed to recognize the errors. The remaining actor generations were accurate. It’s worth noting that, due to the uniform usage of the four exploration principles, out of the 100 actor generations, 50 were either paraphrased questions or questions with new scenarios, these 50 were less prone to errors as they didn’t necessitate significant changes to the mathematical structure. In fact, only 2 solutions among them were incorrect. On the other hand, those that involved restructuring or tackled new topics were more error-prone: 7 from restructuring and 4 from new topics were incorrect. Intriguingly, the critic corrected 6 out of the 7 restructured ones, but only 1 out of the 4 new topics. This might suggest that questions on new topics are more open-ended and might need multiple rounds of interactions for accurate evaluation. Further refinement in this verification process will be the focus of our future work.

Our analysis revealed that the critic excels in evaluation rather than generation. This observation is exemplified in the quoted example, it can be seen that even when the model cannot directly answer the question, it can effectively assess the correctness of the actor’s response. This discovery is intriguing, as the actor’s response effectively serves as a ‘hint’. This observation aligns with prior works that found models are better at evaluating answer (or improving an answer) than generating it from scratch in various settings (Chen et al., 2023; Madaan et al., 2023; Liu et al., 2023; Cobbe et al., 2021, *inter alia*).

[Critic cannot solve question directly]

Input: A magician has 5 hats, each with a different number of balls inside. The first hat has 3 balls, the second hat has twice as many balls as the first, the third hat has half as many balls as the second, the fourth hat has 5 balls, and the fifth hat has 4 times as many balls as the fourth. How many balls are in the magician’s hats altogether?

Output : There are  $3 + 2 * 3 = 7$  balls in the first two hats. There are  $5 - 2 = 3$  balls in the third hat. There are  $4 * 3 = 12$  balls in the fourth and fifth hats. In total, there are  $7 + 3 + 12 = 22$  balls in the magician’s hats. Answer: 22. ✗

[Critic can evaluate actor’s answer]

Input : Please review whether the answer for the following question is correct. Q: A magician has 5 hats, each with a different number of balls inside. The first hat has 3 balls, the second hat has twice as many balls as the first, the third hat has half as many balls as the second, the fourth hat has 5 balls, and the fifth hat has 4 times as many balls as the fourth. How many balls are in the magician’s hats altogether? A: There are  $3 + 2 * 3/2 = 3 + 3 = 6$  balls in the first two hats. There are  $5 - 3/2 = 5 - 1.5 = 3.5$  balls in the third hat. There are  $4 * 5 = 20$  balls in the fourth hat. In total, there are  $6 + 3.5 + 20 = 29.5$  balls. Since you can’t have a fraction of a ball, you would need to round up to the nearest whole number of balls. Therefore, there are 29 balls in the magician’s hats altogether.

Output : The magician has 5 hats, each with a different number of balls inside. The first hat has 3 balls, the second hat has twice as many balls as the first, so it has  $3 * 2 = 6$  balls. The third hat has half as many balls as the second, so it has  $6 / 2 = 3$  balls. The fourth hat has 5 balls. The fifth hat has 4 times as many balls as the fourth, so it has  $5 * 4 = 20$  balls. In total, there are  $3 + 6 + 3 + 5 + 20 = 37$  balls. Since you can’t have a fraction of a ball, you would need to round up to the nearest whole number of balls. Therefore, there are 37 balls in the magician’s hats altogether. The answer provided is incorrect. ✓

## B PROMPT

The critic’s principle is based on the exploration principle but is worded slightly differently from the actor’s exploration principle, in order to tell critic what kind of new questions are diverse and acceptable. The critic will use the principle to evaluate diversity of questions and evaluate the correctness of answer.

### Principles for critique

#### A different topic is acceptable:

Question: Jack is stranded on a desert island. He wants some salt to season his fish. He collects 2 liters of seawater in an old bucket. If the water is 20% salt, how many ml of salt will Jack get when all the water evaporates?

Question (topic): Samantha is designing a circular garden in her backyard. The garden has a diameter of 8 meters. She wants to build a path around the garden that is 1 meter wide. What is the area of the path, in square meters, that Samantha will need to pave with stones or concrete?

#### A question with different structure is acceptable:

Question: Dan owns an ice cream shop and every sixth customer gets a free ice cream cone. Cones cost \$2 each. If he sold \$100 worth of cones, how many free ones did he give away?

Question (restructured): Dan owns an ice cream shop and every sixth customer gets a free ice cream cone. Cones cost \$x each. If he sold \$100 worth of cones, how many free ones did he give away? If we know the answer is 10, what is the value of x?

#### Rephrased question is acceptable:

Question: Joy can read 8 pages of a book in 20 minutes. How many hours will it take her to read 120 pages?

Question (rephrase): How many hours will Joy need to read 120 pages if she can read 8 pages in 20 minutes?

#### A different scenario is acceptable:

Question: Ed has 2 dogs, 3 cats and twice as many fish as cats and dogs combined. How many pets does Ed have in total?

Question (scenario): Sarah owns 4 bicycles, 2 skateboards, and three times as many pairs of rollerblades as bicycles and skateboards combined. How many wheeled sports equipment items does Sarah have in total?

## C EXPERIMENT DETAILS

We use a temperature of 0.7 for the actor during exploration as in prior work (Cobbe et al., 2021), and we sample 10 actor generations for every batch of samples from the replay buffer. We use a temperature of 0.0 for the critic since we found that it performs best. Following prior work (Yuan et al., 2023a), we filter out reasoning paths with incorrect answers or calculations—based on Python evaluation—for the ‘paraphrasing’ and ‘new scenarios’ exploration categories. However, we do not apply this filter to the ‘restructuring’ or ‘new topics’ exploration categories, as we do not have ground truth answers for these two categories. The evaluations for all baselines and our approach are conducted with deterministic sampling following prior work and report  $\text{maj1}@1$  (accuracy) across all experiments. We follow prior work by conducting evaluations using deterministic sampling for both

our approach and the baseline methods. We report maj1@1 accuracy across all experimental setups. All models are trained with the same hyperparameters: global batch size = 128, learning rate =  $2e-5$ , epochs = 3, sequence length = 2048. The training is done with 8x A100 80GB GPUs.