G Detailed Experimental Settings

All our experiments are performed on 8 × A100 GPUs and finished in about one day. We enable the actor-learner collocation supported by Oat (Liu et al., 2025a) to optimize the training efficiency. We show the experimental configurations in Table 6.

Parameter	Value
ACTOR	
Maximum response length	3000 tokens
Sampling temperature	1.0
(top P, top k)	(1.0, -1)
Number of responses per question	8
LEARNER	
Optimizer	AdamW
Adam parameters (β_1, β_2)	(0.9, 0.95)
Weight decay	0.0
Gradient norm clipping	1.0
Learning rate scheduler	Constant
Learning rate	1×10^{-6}
Inner proximal update epoch	1
KL loss coefficient	0.0
KL penalty coefficient	0.0
Policy clipping parameter	0.2

Table 6: Hyperparameter configurations used in all experiments.

H Prompts Used for GPT-As-A-Judge

Prompt for checking the model's question-answering ability.

Prompt for Checking Question-Answering Ability

I will send you a question and a long response generated by an LLM. Your task is to determine whether the output attempts to answer the question or not. The output may sometimes include irrelevant content, hallucinations, or random, off-topic responses.
Please classify the output into one of the following categories:
Output Format:
Your response must start with a **single integer** (0 or 1), followed by a **brief explanation**.

- **Return 0:** → The output is not trying to answer the question (e.g., irrelevant content, random talking, hallucinations). *Example output:* '0: The response is off-topic and does not address the question.'
- **Return 1:** → The output attempts to answer the question, regardless of how complete or accurate the answer is. *Example output:* '1: The response engages with the question, even if the answer is incomplete or incorrect.'

Question: {question}
Response: {response}

Prompt for LLM-based detection to determine whether a response contains self-reflection behaviors.

LLM-based Detection for Self-Reflection

I will send you a mathematical question along with a detailed response. Your task is to determine whether the response is attempting to answer the question. If the response is off-topic, hallucinated, random talk, or otherwise irrelevant, mark it as **0**. Otherwise, assess whether the response exhibits self-reflection.
Categorization Rules:

1. **Category 0**: The response is **off-topic, nonsensical, incoherent, overly repetitive, or lacks logical reasoning**.
   - Example cases:
     - The response does not relate to the question.
     - It contains meaningless or hallucinated content.
     - It consists of excessive repetition without coherence.
2. **Category 1**: The response **attempts to answer the question** but does **not** exhibit self-reflection.
   - Example cases:
     - The response directly solves the problem without revisiting steps.
     - No attempt is made to verify the correctness of the answer or explore alternative solutions.
3. **Category 2**: The response **demonstrates self-reflection** at any level.
   - This may include:
     - **Explicit self-reflection keywords**, such as: *recheck, rethink, reassess, reevaluate, re-evaluate, reevaluation, re-examine, reexamine, reconsider, reanalyze, double-check, check again, think again, verify again, go over the steps*, etc.
     - **Implicit self-reflection behaviors**, such as revisiting the solution, questioning assumptions, or considering alternative approaches **without explicit keywords**.
   - If any form of self-reflection is present, **always categorize it as 2**, regardless of correctness or answer quality.
4. **Category 3**: The response consists **solely of Python code for calculations** without exhibiting self-reflection.
   - Example cases:
     - The response only provides a Python script to compute the solution **without any verification, re-evaluation, or alternative considerations**.

Output Format:
Your response should first provide a **very brief explanation** of your analysis, followed by a **single category number (0, 1, 2, or 3)** at the end. You must include the category number at the end of your response.
Example outputs:

- 'The response is off-topic and does not attempt to answer the question. 0.'
- 'The response provides a direct solution without self-reflection. 1.'
- 'The response demonstrates self-reflection. 2.'
- 'The response consists solely of Python code without any self-reflection. 3.'

Question: {question}
Response: {response}