and "try again" frequently result in false positive detections. To reduce false positives, we maintain a small, highly selective keyword pool consisting of terms that are strongly indicative of self-reflection. In our experiment, the keyword pool is limited to: recheck, rethink, reassess, reevaluate, re-evaluate, reevaluation, re-examine, reexamine, reconsider, reanalyze, double-check, check again, think again, verify again, and go over the steps.

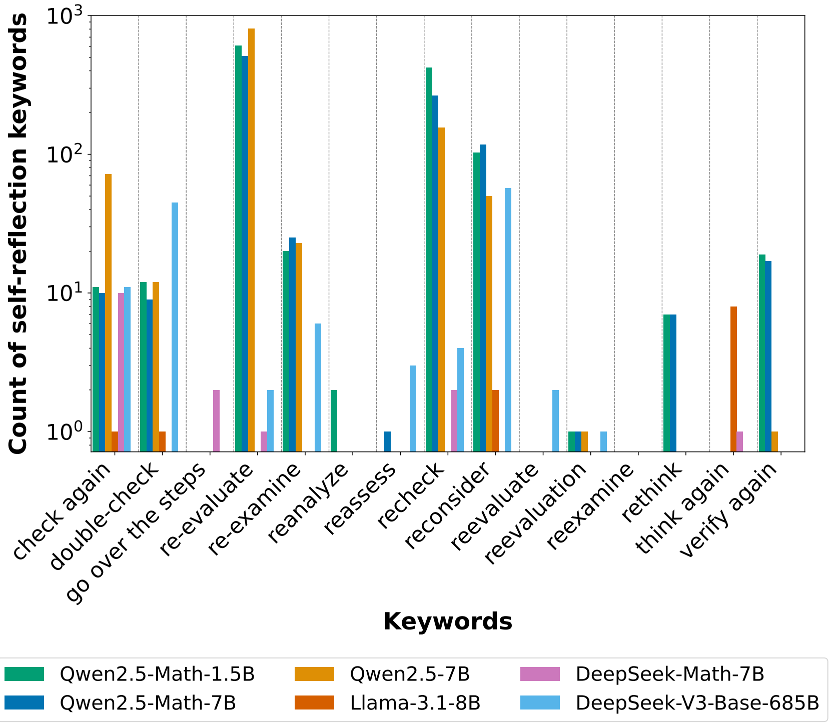

Figure 10: Count of keyword occurrences out of 40,000 responses (500 questions × 8 responses per question × 10 temperatures). y is in log scale.

We present the occurrences of various keywords in the responses generated by different models in Figure 10. Interestingly, different model families emphasize different keywords. For instance, phrases such as "check again", "double-check", "re-evaluate", "re-examine", "recheck", "reconsider", and "verify again" appear most frequently in the Qwen2.5 family. In contrast, "re-evaluate", "re-examine", and "verify again" do not appear in the responses of the DeepSeek family, while Llama models frequently use the phrase "think again." We hypothesize that this phenomenon results from differences in the pretraining data, particularly in relation to reasoning and mathematics.

Although we meticulously select the keyword pool, it may still be insufficient to identify some implicit behaviors of self-reflection that do not contain a specific keyword. Additionally, it can lead to false positives, as illustrated in Case (a) of Figure 11. To address these limitations and more accurately assess the self-reflection capability of base models, we leverage stronger LLMs (`GPT-4o-mini` in our experiments) to analyze the responses and determine whether they exhibit explicit self-reflection (e.g., keywords like "recheck" and "reevaluate") or implicit self-reflection (e.g., more sophisticated patterns that cannot be easily captured through keyword matching). This approach helps distinguish true self-reflection behaviors from superficial or incidental use of related terms.

While LLM-based detection effectively filters out false positives from keyword-based detection and identifies implicit self-reflection behaviors, it can still misclassify responses, particularly when they are lengthy and complex. For instance, Case (b) in Figure 11 shows a false positive in LLM-based detection, where the response is categorized as self-reflection by the LLM but does not actually exhibit self-reflection. This type of error can be filtered out by keyword-based detection. To enhance robustness, we integrate keyword-based

Case(a): What is the smallest positive multiple of 23 that is 4 more than a multiple of 89?

Response

To find the smallest positive multiple of 23 that, ...
Start with the smallest possible value of k by checking if they satisfy the equation. If not, increase k and check again.
...
Now, trying $k = 21$: $23 \cdot 21 = 483$ ...
...

Case(b): What is the greatest multiple of 3 that is less than 500 and is also a multiple of 4?

Response

To find the greatest multiple of 3 that is less than 500 and is also a multiple of 4, I need to first consider the greatest multiple of 4 less than 500, which is 496. However, 496 is not a multiple of 3. I then need to consider the next multiple of 4, which is 492. 492 is a multiple of 3, since 492 / 3 = 164 with no remainder.</think>
<answer> 492 <answer>

Figure 11: **Case (a)**: a false positive in keyword-based detection. **Case (b)**: a false positive in LLM-based detection.

and LLM-based detection through cross-validation. The combined detection results, along with the individual results from keyword-based and LLM-based methods, are presented in Figure 12.

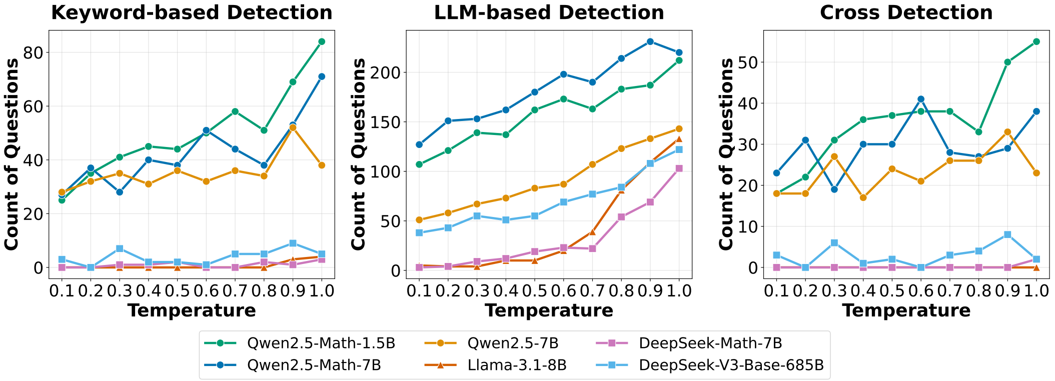

Figure 12: Comparison of keyword-based detection, LLM-based detection, and cross detection. Self-reflections are counted at the question level across 500 questions, where a question is marked as having self-reflection if at least one of its eight responses exhibits self-reflection.

E Examples of Aha Moment in DeepSeek-V3-Base

Fig. 13 shows two examples to demonstrate that the DeepSeek-V3-Base model already exhibits the so-called "aha moment" even before the RL-tuning.

F Comparison Between DeepSeek-V3-Base and DeepSeek-R1-Zero

We analyze DeepSeek-V3-Base and DeepSeek-R1-Zero to understand changes in model behavior during R1-Zero training. In Fig. 14, we present the breakdown of response categories across difficulty levels for 500 MATH questions evaluated on both models. The results indicate that most incorrect responses are corrected after RL training, demonstrating substantial performance gains from R1-Zero training. Meanwhile, we find an increase in unformatted responses, which aligns with the observation in Liu et al. (2025b).

In Table 5, we report the average response lengths across categories. Note that truncated responses would fall into any of the other three categories if a larger context size were used; thus, we exclude them from the table. The results show a substantial increase in response lengths across all categories, including correct responses, consistent with the results in the

Question(a): An element is randomly chosen from among the first 15 rows ...	**Question(b):** Find the least positive integer n such that $1/\sin 45^\circ \sin 46^\circ + ...$
Response In Pascal's Triangle, every row starts and ends with 1, ... This can be calculated as: *awkward silence* Wait, I'm overthinking. Let's try again. The number of elements in the first n rows of Pascal's Triangle is $(n+1)!/[2! * (n-1)!]$. ...	**Response** First, I notice that the left-hand side of the equation has the form ... Simplifying, I get ... Aha! I can use this to get $\sin(a+2b) = \sin((a+b)+b) = 3\sin(a+b)$ ... This looks messy, but I notice that it has the form $\sin a \cos b + \cos a \sin b$ again. This gives me an idea. Maybe I can ... ...

Figure 13: Cases showing that DeepSeek-V3-Base already exhibits "Aha moment" even before RL tunning.

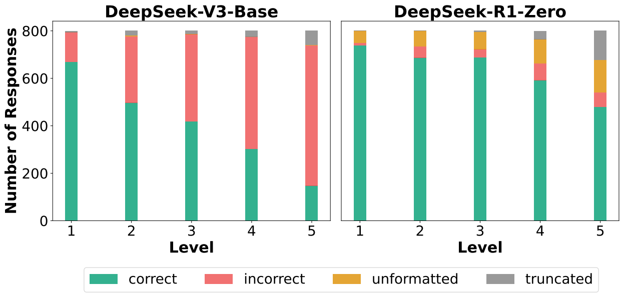

Figure 14: Breakdown of response categories across difficulty levels in the MATH dataset for DeepSeek-V3-Base and DeepSeek-R1-Zero.

Category	Base	R1-Zero
Unformatted	880.7	7870.3
Correct	621.3	4965.4
Incorrect	1038.9	8206.1

Table 5: Average response string lengths across categories for DeepSeek-V3-Base (Base) and DeepSeek-R1-Zero (R1-Zero).

Fig. 3 of Guo et al. (2025). However, the average length of incorrect responses is notably longer than that of correct responses. We hypothesize this is because more challenging questions generally require longer responses due to increased reasoning complexity, and incorrect responses are more likely to originate from harder questions, resulting in a longer average length.

Self-reflection does not necessarily imply higher accuracy. To investigate whether self-reflection behaviors are associated with model performance during the inference (acknowledging that self-reflection may improve exploration during training—a potential positive effect outside this section's scope), we analyze questions that elicit at least one response with self-reflection from DeepSeek-R1-Zero across eight trials. For each question, we sample 100 responses and divide them into two groups: those with self-reflection and those without. We then compute the accuracy difference between these two groups for each question. As shown in Fig. 15, the results indicate that nearly half responses with self-reflection do not achieve higher accuracy than those without self-reflection, suggesting that self-reflection does not necessarily imply higher inference-stage accuracy for DeepSeek-R1-Zero.

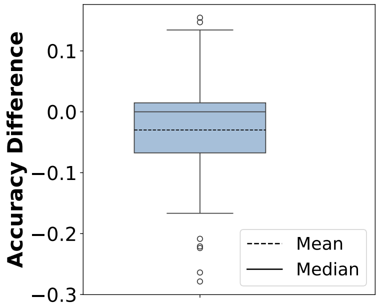

Figure 15: Accuracy difference between responses with and without self-reflection for each question (responses sampled from DeepSeek-R1-Zero).