A. Implementation Details

A.1. RLVR Algorithms

To reduce memory and computational overhead, several critic-free variants have been proposed. GRPO (Shao et al., 2024) estimates the advantage with a normalized reward within a group of responses to the same question: $A_i = [r_i - \text{mean}(\mathbf{r})]/\text{std}(\mathbf{r})$, where $\mathbf{r} = \{r_1, \ldots, r_G\}$ denotes the set of rewards for a group of G sampled responses. RLOO (Ahmadian et al., 2024) instead adopts a leave-one-out baseline within each batch $\mathcal{B}$. Its advantage is defined as $A_i = r_i - \frac{1}{|\mathcal{B}|-1}\sum_{j\neq i} r_j$.

A.2. Low-Variance *pass@k* Estimation

Directly computing pass@k using only k sampled outputs per problem can lead to high variance. To mitigate this, we follow the unbiased estimation method proposed by Chen *et al.* (Chen et al., 2021). Specifically, for each problem x_i from the evaluation dataset $\mathcal{D}$, we generate n samples ($n \geq k$) and count the number of correct samples as c_i. The unbiased estimator of pass@k over the dataset is given by:

$$\text{pass@}k := \mathbb{E}_{x_i \sim \mathcal{D}}\left[1 - \frac{\binom{n-c_i}{k}}{\binom{n}{k}}\right] \tag{2}$$

With this formulation, we can easily estimate pass@k with low variance across all $k \leq n$.

In our experiments, we set n to the largest (*i.e.*, rightmost) k value in the pass@k curves, typically 128, 256, or 1024. For example, in Figure 2, we use $n = 128$ for MATH500, Minerva, and GSM8K, and $n = 1024$ for AMC23 and AIME24. For the Olympiad benchmark, we set $n = 128$ for the Qwen models and $n = 1024$ for LLaMA-3.1-8B, due to its relatively lower base model capacity.

B. More Related Works

Reinforcement Learning for LLM Reasoning. Since the emergence of LLMs, the post-training phase has proven crucial to enhance problem solving and reasoning abilities (Ouyang et al., 2022). This stage typically falls into three main categories: supervised fine-tuning using human-curated or distilled data (Wang et al., 2023), self-improvement iteration (Zelikman et al., 2022; Gulcehre et al., 2023), and reinforcement learning (Ouyang et al., 2022). Previously, a reward model or preferences between responses were employed for reward modeling (Ouyang et al., 2022; Rafailov et al., 2023). Recently, Reinforcement Learning with Verifiable Rewards (RLVR) has gained significant traction as a method to improve the reasoning capabilities of LLMs in domains such as mathematics and programming (Lambert et al., 2024; Shao et al., 2024). An encouraging landmark work is OpenAI's o1 model (Jaech et al., 2024), which was among the first large-scale applications of RL for reasoning, achieving state-of-the-art results at the time of its release. Following this, Deepseek-R1 (Guo et al., 2025) became the first open-weight model to match or surpass the performance of o1. A significant innovation introduced with R1 is the "zero" setting, where reinforcement learning is applied directly to the base LLM, bypassing any intermediate supervised tuning. This approach inspired a wave of open-source efforts to replicate or extend R1's methodology and improve RL algorithms (Zeng et al., 2025; Liu et al., 2025b; Yu et al., 2025; Liu & Zhang, 2025; Zhao et al., 2025a; Wang et al., 2025). In parallel, reinforcement learning has also gained attention in the multimodal domain, driving advancements in multimodal reasoning (Chen et al., 2025a; Shen et al., 2025; Zheng et al., 2025).

Analysis of RLVR. Although there are many excellent open-source works and algorithmic designs in the field of RLVR, there remains a lack of deep understanding regarding the root effects of RLVR on LLM reasoning abilities and its limitations when starting from the base model. Several studies (Liu et al., 2025a; Zhao et al., 2025b; Shah et al., 2025) highlight that the reflective behaviors observed in R1-like models actually emerge from the base models, rather than being introduced by RLVR training. Dang *et al.* (Dang et al., 2025) observed a phenomenon similar to our findings: Pass@k deteriorates rapidly and fails to recover with reinforcement learning, but this was seen only in a limited experimental setup with Qwen-2.5-0.5B on GSM8K. More importantly, they did not explore the relationship between the base model and the RL model. In contrast, our paper conducts systematic and rigorous experiments to show that not only reflective behaviors but all reasoning paths are already embedded in the base model. We further demonstrate that RLVR does not elicit any new reasoning abilities beyond the base model.

C. Detailed Experimental Results

C.1. More Results on Mathematics and Coding

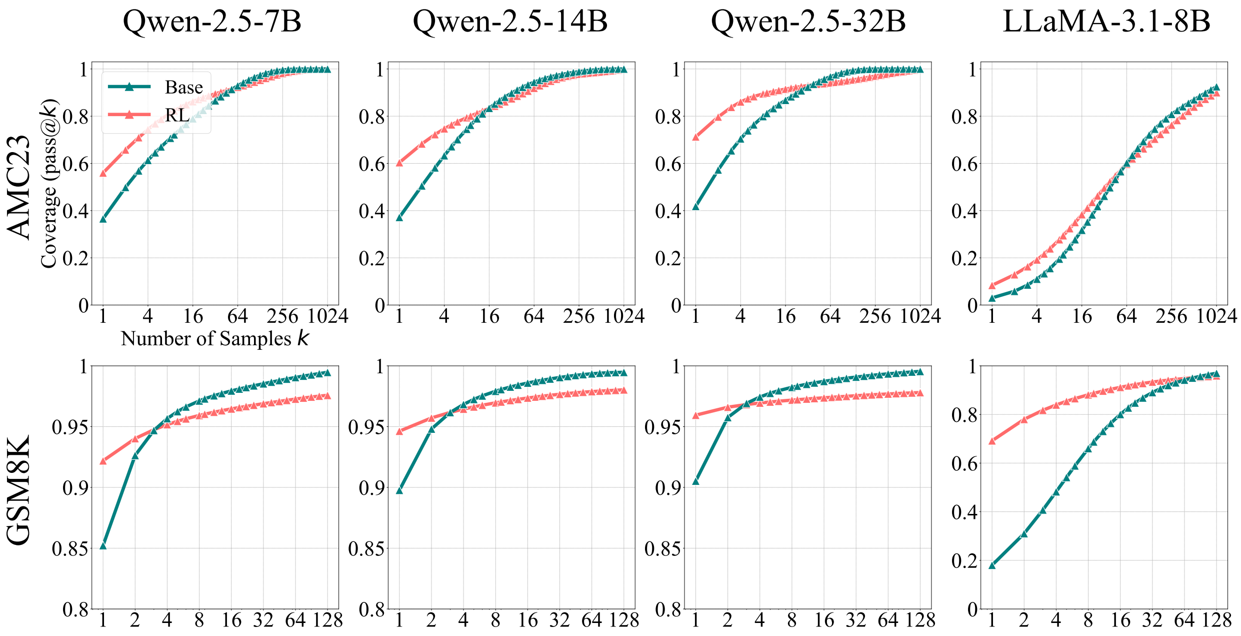

Figure 10: More results of SimpleRLZoo on GSM8K and AMC23.

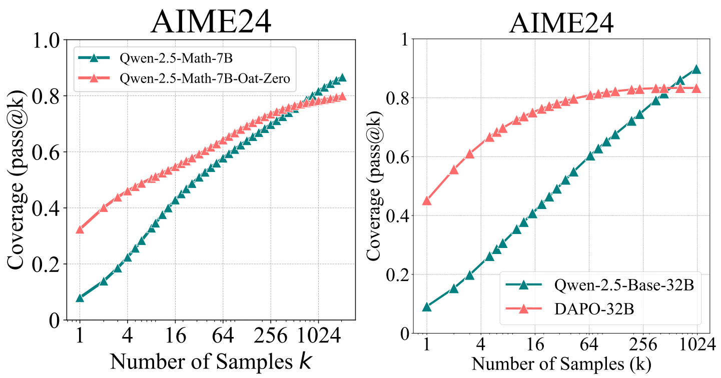

Figure 11: Oat-Zero-7B and DAPO-32B are evaluated on AIME24 and compared against their respective base models.

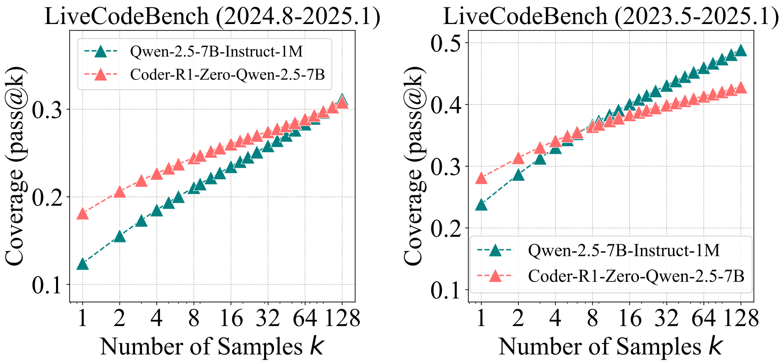

Figure 12: Coder-R1 on LiveCodeBench.

C.2. Validity of Chain-of-Thought on AIME24

We manually check the CoTs for the most challenging AIME24 benchmark. To begin, we introduce a filtering mechanism designed to eliminate easily guessable problems. Specifically, we prompt Qwen2.5-7B-Base to answer questions directly, without using chain-of-thought reasoning, and sample answers multiple times. If a problem can be answered correctly with a low but non-zero probability (e.g., <5%), we consider it to be guessable and remove it. Problems that can be directly answered correctly with a high probability are retained, as they are likely easier and solvable using valid CoTs.

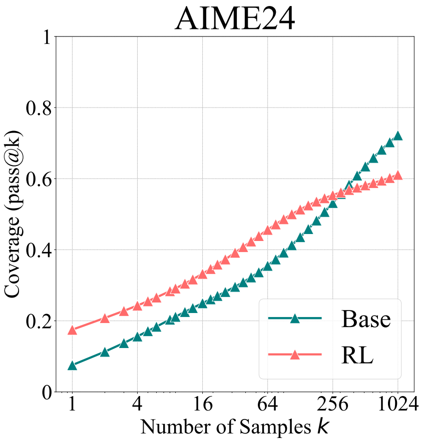

Figure 13: Pass@k curves of the base and SimpleRLZoo-7B models in the filtered AIME24.

The base and RL model pass@k curves on this filtered AIME24 can be found in Figure 13, showing a similar trending to previous results. Although this filtering method is heuristic, it proves to be effective. Applying it to AIME24 (30 questions) results in a subset of 18 questions. We then prompt the models to answer these filtered questions using CoT reasoning. Then we perform a manual inspection of all CoTs that led to correct answers on the hardest problems – those with an average accuracy below 5%. The base model answered 7 such questions, with 5 out of 6 containing *at least one* correct CoT (excluding one ambiguous case of correctness due to skipped reasoning steps). Similarly, the RL-trained model answered 6 questions, 4 of which included *at least one* correct CoT. These results suggest that even for the hardest questions in the challenging AIME24, base model can sample valid reasoning paths to solve the problems.