C.3. Accuracy Distribution Visulization

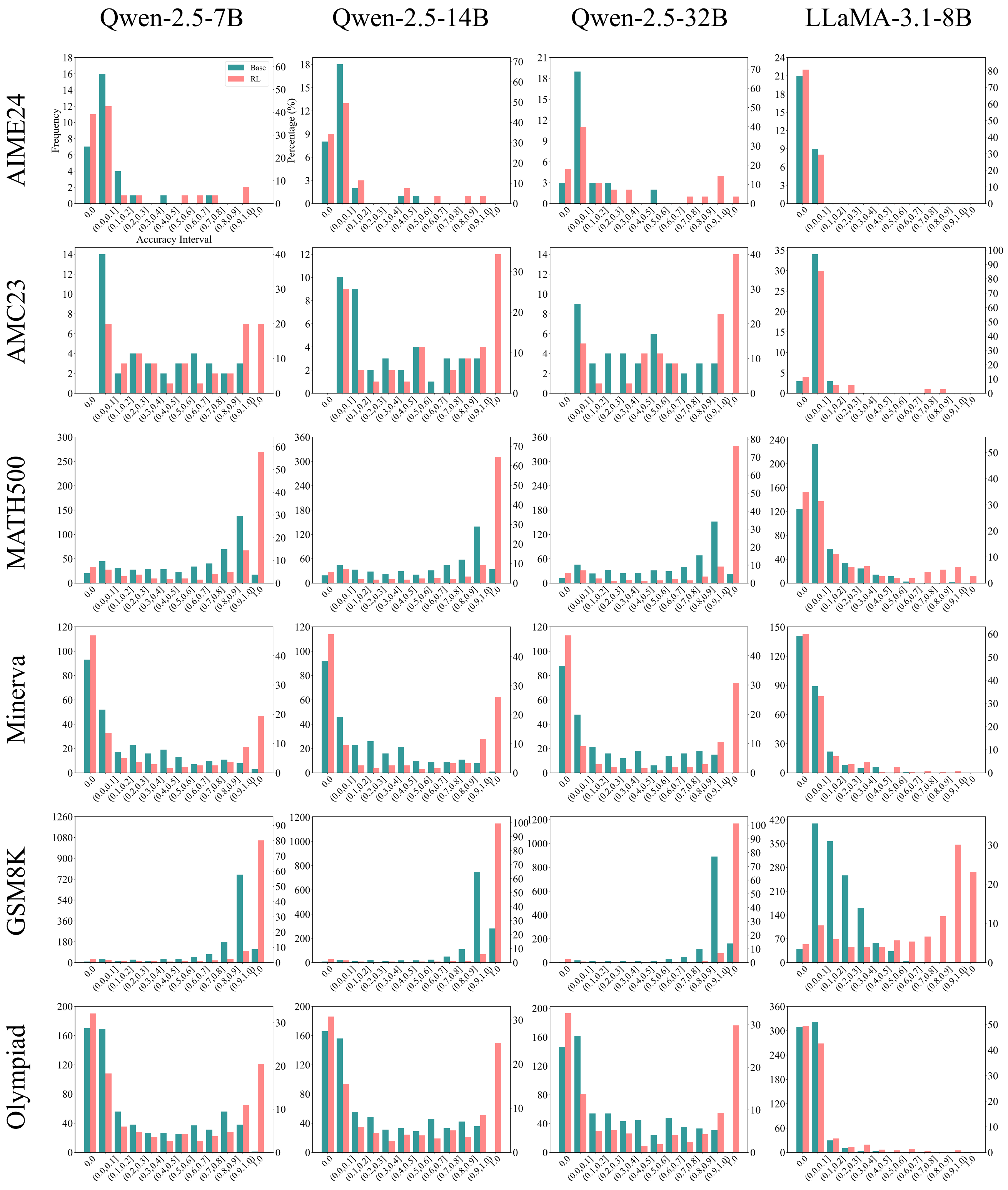

Figure 14: Accuracy histogram before and after RLVR with SimpleRLZoo models.

C.4. Perplexity Analysis

To analyze how perplexity evolves over the course of RLVR training, we evaluated three RLVR checkpoints–early, middle, and final in Section 4.3 RL training. For each checkpoint, we sampled 32 responses per problem, computed the median among 32 perplexity values, and reported the average over the first 10 problems in the table. As expected, we observed that $\mathrm{PPL}_{\mathrm{Base}}(\mathbf{Y}_{\mathrm{RL}}|x)$ gradually decreases as RL training progresses, indicating that RLVR mainly sharpens the distribution within the base model's prior rather than expanding beyond it.

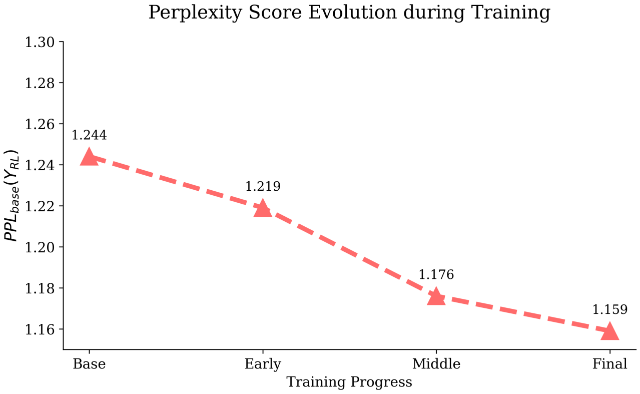

Figure 15: Perplexity Evolution during RL Training.

C.5. Different RLVR Algorithms

We report several additional observations on different RLVR algorithms in Figure 8. First, DAPO achieves slightly higher pass@1 scores across all three datasets; however, its dynamic sampling strategy requires approximately 3 $\sim$ 6$\times$ more samples per batch during training compared to other algorithms. Moreover, its performance drops significantly at $k = 256$. Second, RLOO and Reinforce++ perform consistently well across the entire k range (from 1 to 256), while maintaining efficient training costs, achieving a good balance between effectiveness and efficiency. Third, ReMax shows lower performance at both pass@1 and pass@256. We hypothesize that this is due to its use of the greedy response reward as the advantage baseline, which in the RLVR setting is binary (0 or 1) and highly variable. This likely results in unstable gradient updates during training.

Table 3: Detailed values for each point at pass@1 and pass@256 across different RL algorithms in Figure 8.

Model	**Omni-MATH-Train**		**Omni-MATH-Test**		**MATH500**	
	pass@1	pass@256	pass@1	pass@256	pass@1	pass@256
Qwen2.5-7B	9.9	67.2	10.2	69.1	34.5	96.2
GRPO	26.1	66.3	25.1	68.3	74.4	97.2
PPO	27.2	65.8	26.8	69.2	75.2	97.2
ReMax	24.4	65.5	23.8	67.5	73.5	96.6
RLOO	28.6	66.4	**28.1**	69.2	75.0	**97.4**
Reinforce++	28.2	**67.7**	**28.0**	**69.7**	75.4	96.8
DAPO	**31.4**	66.1	26.5	67.0	**75.6**	96.4

Table 4: Detailed values at pass@1 and pass@256 across different RL training steps in Figure 1 (right).

Model	**Omni-MATH-Train**		**Omni-MATH-Test**		**MATH500**	
	pass@1	pass@256	pass@1	pass@256	pass@1	pass@256
Qwen2.5-7B	9.9	**67.2**	10.2	**69.1**	34.5	96.2
GRPO-step150	26.1	66.3	25.1	68.3	74.4	**97.2**
GRPO-step300	33.6	65.3	27.1	66.6	75.4	96.0
GRPO-step450	**42.5**	64.3	**28.3**	63.9	**76.3**	95.4

C.6. Effects of KL and Rollout Number

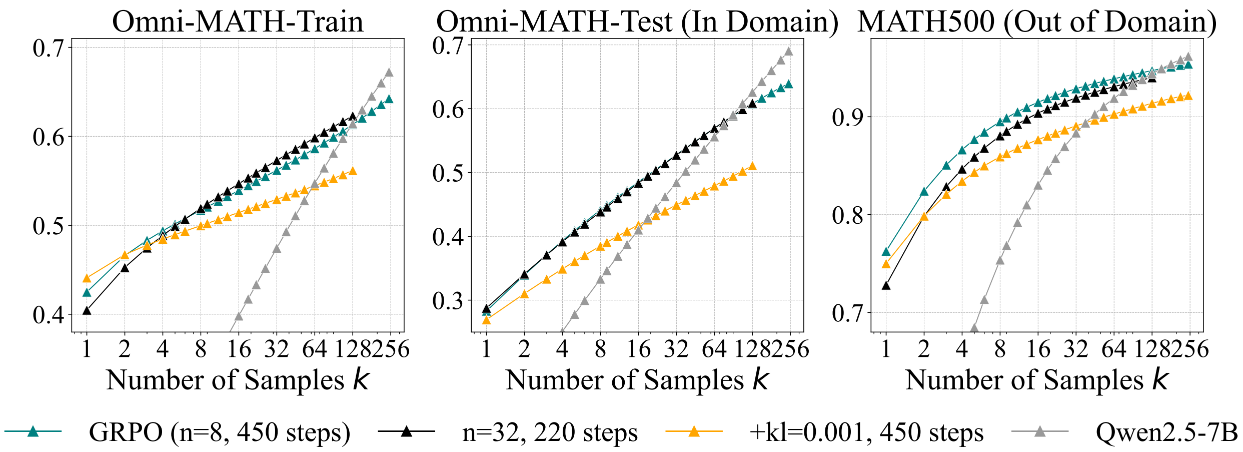

Figure 16: **Ablation Study on KL Loss and Rollout Number** n**.** For increasing n from 8 to 32, we keep the prompt batch size unchanged, which results in increased computation per training step. Due to resource constraints, we train for only 220 steps under this setting, leading to lower pass@1 as the model has not yet converged. Nevertheless, the model with $n = 32$ achieves a higher pass@128, highlighting the positive effect of larger rollout numbers in improving pass@k at higher values of k.

C.7. Solvable Problem Coverage Analysis

Table 2 reports the fraction of problems categorized as four conditions: (1) both models solve the problem at least once, (2) only the base model solves it, (3) only the RLVR model solves it, and (4) neither model solves it in any of the k samples. It highlights that there are many cases where the base model solves a problem but RLVR fails (type 2), and very few where RLVR succeeds while the base does not (type 3). Even in the rare type 3 cases (e.g., 1% or about 5 problems in MATH500), the base model is able to solve all of them when sampling 1024 times. These results support our conclusion that RLVR rarely solves problems the base model cannot and generally results in reduced coverage.

Table 5: Indices of solvable problems in AIME24 (starting from 0). An approximate subset relationship can be observed: most problems solved by the RL model are also solvable by the base model.

Models	**Problem Indices**
Qwen2.5-7B-Base	0, 1, 4, 6, 7, 8, 9, 11, 12, 14, 15, 16, 17, 18, 19, 22, 23, 24, 25, 26, 27, 28, 29
SimpleRL-Qwen2.5-7B	0, 1, 6, 7, 8, 9, 12, 14, 15, 16, 18, 22, 23, 24, 25, 26, 27, 28, 29

Table 6: Indices of solvable problems in LiveCodeBench (ranging from 400 to 450, starting from 0).

Model	**Solvable Problem Indices**
Qwen2.5-7B-Instruct-1M	400, 402, 403, 407, 409, 412, 413, 417, 418, 419, 422, 423, 427, 432, 433, 436, 438, 439, 440, 444, 445, 448, 449
Coder-R1	400, 402, 403, 407, 412, 413, 417, 418, 419, 422, 423, 427, 430, 433, 438, 439, 440, 444, 445, 449