this comparison. The results, summarized in Table 21, show that distillation achieves significantly better performance than reinforcement learning while requiring approximately only 1/10 of the GPU hours. Furthermore, distillation from teacher logits enables the student model to expand its exploration space and enhance its reasoning potential, as evidenced by the improved pass@64 scores on the AIME'24 and AIME'25 benchmarks after distillation, compared to the initial checkpoint. In contrast, reinforcement learning does not lead to any improvement in pass@64 scores. These observations highlight the advantages of leveraging a stronger teacher model in guiding student model learning.

Table 21: Comparison of reinforcement learning and on-policy distillation on Qwen3-8B. Numbers in parentheses indicate pass@64 scores.

Method	AIME'24	AIME'25	MATH500	LiveCodeBench v5	MMLU -Redux	GPQA -Diamond	GPU Hours
Off-policy Distillation	55.0 (90.0)	42.8 (83.3)	92.4	42.0	86.4	55.6	-
+ Reinforcement Learning	67.6 (90.0)	55.5 (83.3)	94.8	52.9	86.9	61.3	17,920
+ On-policy Distillation	**74.4 (93.3)**	**65.5 (86.7)**	**97.0**	**60.3**	**88.3**	**63.3**	1,800

The Effects of Thinking Mode Fusion and General RL To evaluate the effectiveness of Thinking Mode Fusion and General Reinforcement Learning (RL) during the post-training, we conduct evaluations on various stages of the Qwen-32B model. In addition to the datasets mentioned earlier, we introduce several in-house benchmarks to monitor other capabilities. These benchmarks include:

- **CounterFactQA**: Contains counterfactual questions where the model needs to identify that the questions are not factual and avoid generating hallucinatory answers.
- **LengthCtrl**: Includes creative writing tasks with length requirements; the final score is based on the difference between the generated content length and the target length.
- **ThinkFollow**: Involves multi-turn dialogues with randomly inserted `/think` and `/no_think` flags to test whether the model can correctly switch thinking modes based on user queries.
- **ToolUse**: Evaluates the stability of the model in single-turn, multi-turn, and multi-step tool calling processes. The score includes accuracy in intent recognition, format accuracy, and parameter accuracy during the tool calling process.

Table 22: Performance of Qwen3-32B after Reasoning RL (Stage 2), Thinking Mode Fusion (Stage 3), and General RL (Stage 4). Benchmarks with * are in-house datasets.

		Stage 2 Reasoning RL	Stage 3 Thinking Mode Fusion		Stage 4 General RL	
	Benchmark	Thinking	Thinking	Non-Thinking	Thinking	Non-Thinking
General Tasks	LiveBench 2024-11-25	68.6	70.9+2.3	57.1	74.9+4.0	59.8+2.8
	Arena-Hard	86.8	89.4+2.6	88.5	93.8+4.4	92.8+4.3
	CounterFactQA*	50.4	61.3+10.9	64.3	68.1+6.8	66.4+2.1
Instruction & Format Following	IFEval strict prompt	73.0	78.4+5.4	78.4	85.0+6.6	83.2+4.8
	Multi-IF	61.4	64.6+3.2	65.2	73.0+8.4	70.7+5.5
	LengthCtrl*	62.6	70.6+8.0	84.9	73.5+2.9	87.3+2.4
	ThinkFollow*	-	88.7		98.9+10.2	
Agent	BFCL v3	69.0	68.4-0.6	61.5	70.3+1.9	63.0+1.5
	ToolUse*	63.3	70.4+7.1	73.2	85.5+15.1	86.5+13.3
Knowledge & STEM	MMLU-Redux	91.4	91.0-0.4	86.7	90.9-0.1	85.7-1.0
	GPQA-Diamond	68.8	69.0+0.2	50.4	68.4-0.6	54.6+4.3
Math & Coding	AIME'24	83.8	81.9-1.9	28.5	81.4-0.5	31.0+2.5
	LiveCodeBench v5	68.4	67.2-1.2	31.1	65.7-1.5	31.3+0.2

The results are shown in Table 22, where we can draw the following conclusions:

(1) Stage 3 integrates the non-thinking mode into the model, which already possesses thinking capabilities after the first two stages of training. The ThinkFollow benchmark score of 88.7 indicates that the model has developed an initial ability to switch between modes, though it still occasionally makes errors. Stage 3 also enhances the model's general and instruction-following capabilities in thinking mode, with CounterFactQA improving by 10.9 points and LengthCtrl by 8.0 points.