Multi-Turn Effective Answer Ratio Before and After RL Training

Legend: Before RL, After RL

Method and Model Size	Before RL	After RL
DAPO 32B	88.7	63.3
Dr. GRPO 7B	57.5	40.4
Dr. GRPO 1.5B	43.8	19.6
PPO 32B	88.7	9.6
PPO 7B	79.4	9.1
PPO 1.5B	38.4	31.4
PPO 0.5B	34.9	24.0
PPO + UFO 3B (Ours)	79.7	92.8

Y-axis: Unique Answer Ratio (%)