NeurIPS 2026

DUET

Optimize Token-Budget Allocation for Reinforcement Learning with Verifiable Rewards

Haoyu Hu[1] Xuandong Zhao[2] Xuhai "Orson" Xu[3] Nori Jacoby[1]

[1] Cornell University [2] UC Berkeley [3] Columbia University

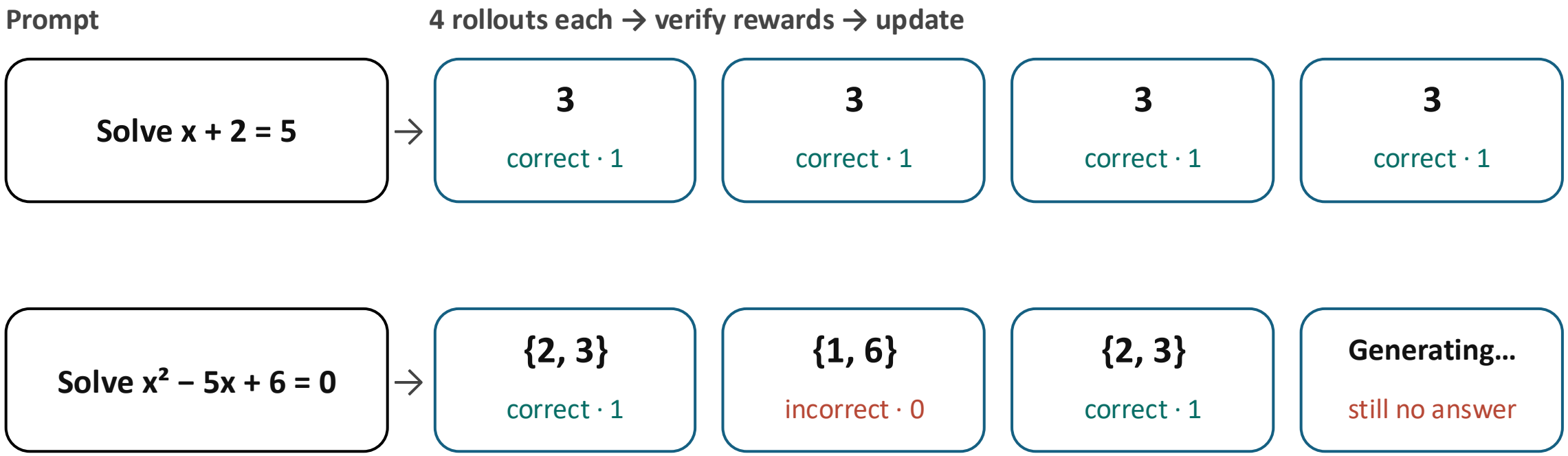
Faster training with higher quality

Paper

Code

Where standard RLVR wastes compute

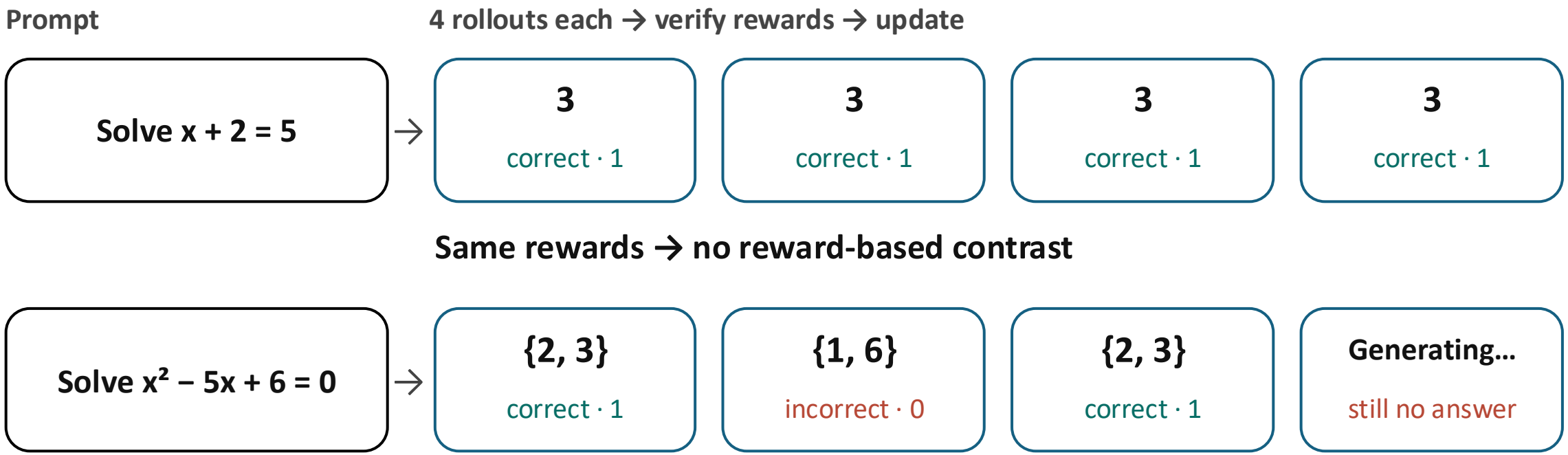
1 Sample answers and verify their rewards



Generation is paid for before we know how useful each attempt is.

Where standard RLVR wastes compute

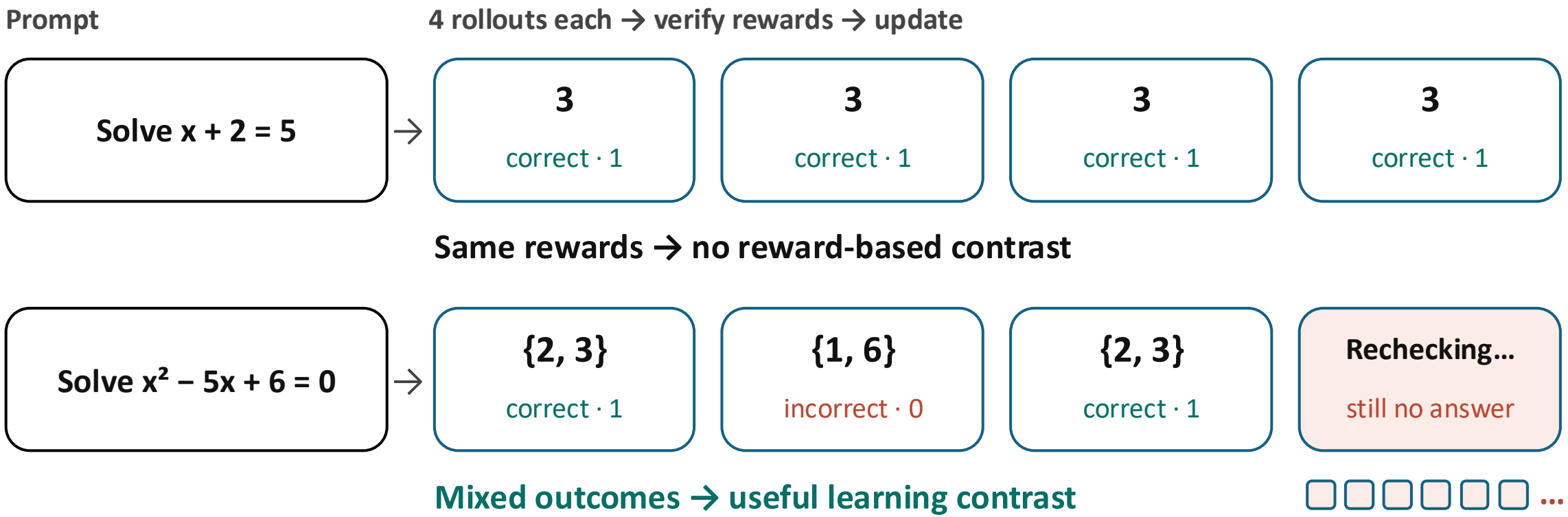
2 Equal rollout counts do not mean equal learning value



Fixed counts spend equally on prompts with different learning value.

Where standard RLVR wastes compute

3 A stalled rollout can keep the whole update waiting

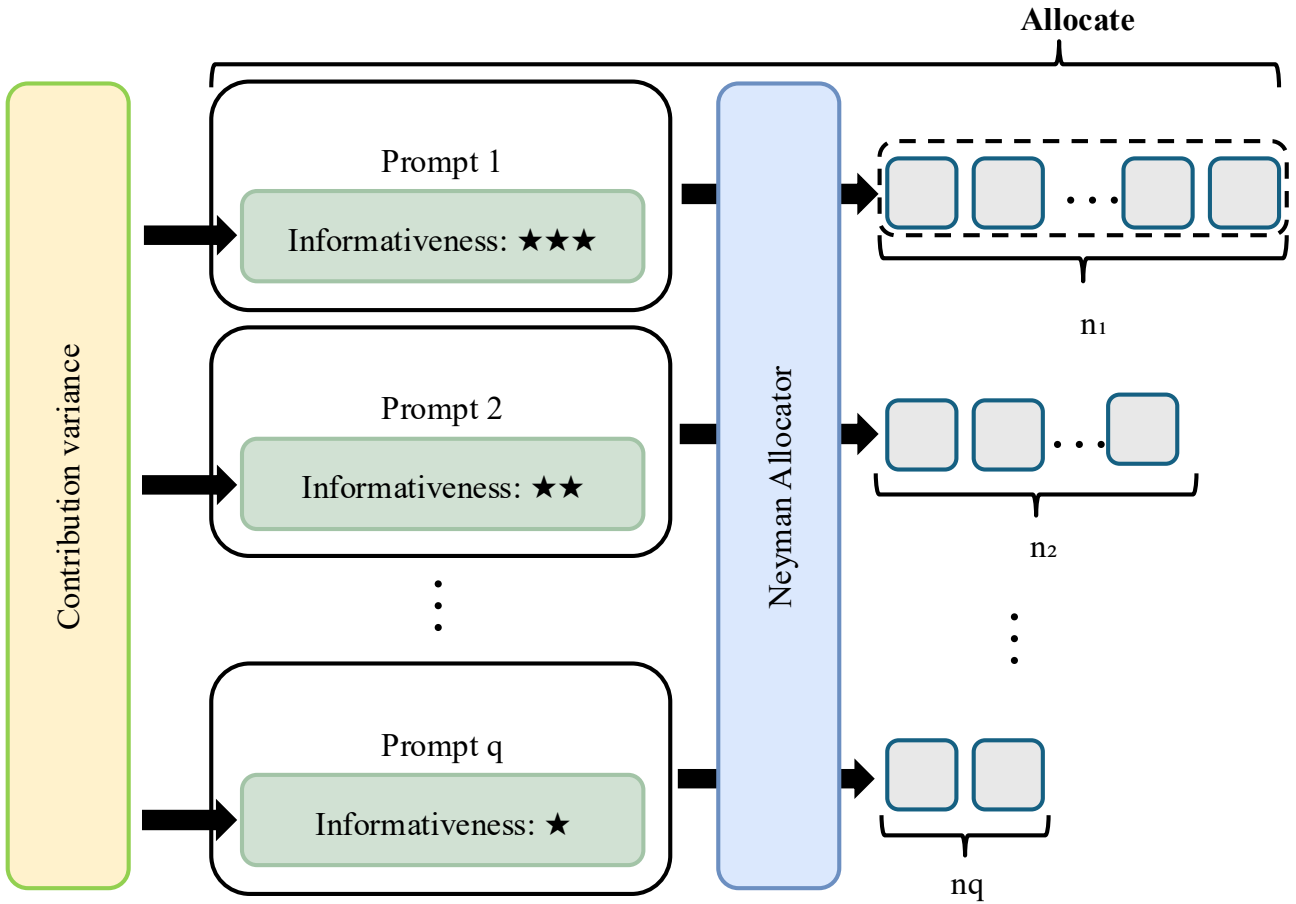


Stalled decoding spends more tokens and delays the next update.

Where should we spend rollouts—and when should we stop?

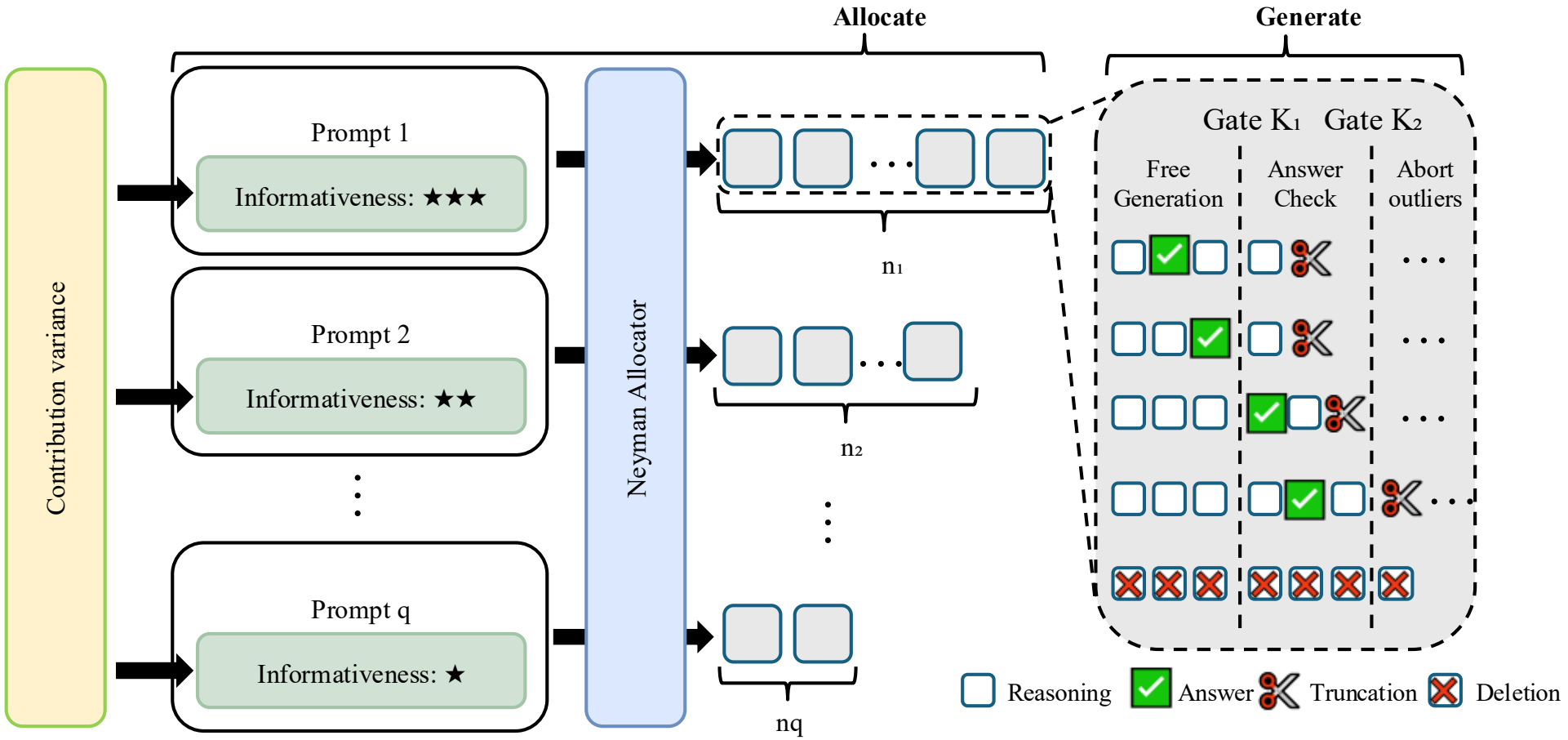
DUET: allocate → generate → update

1 Allocate more rollouts to informative prompts



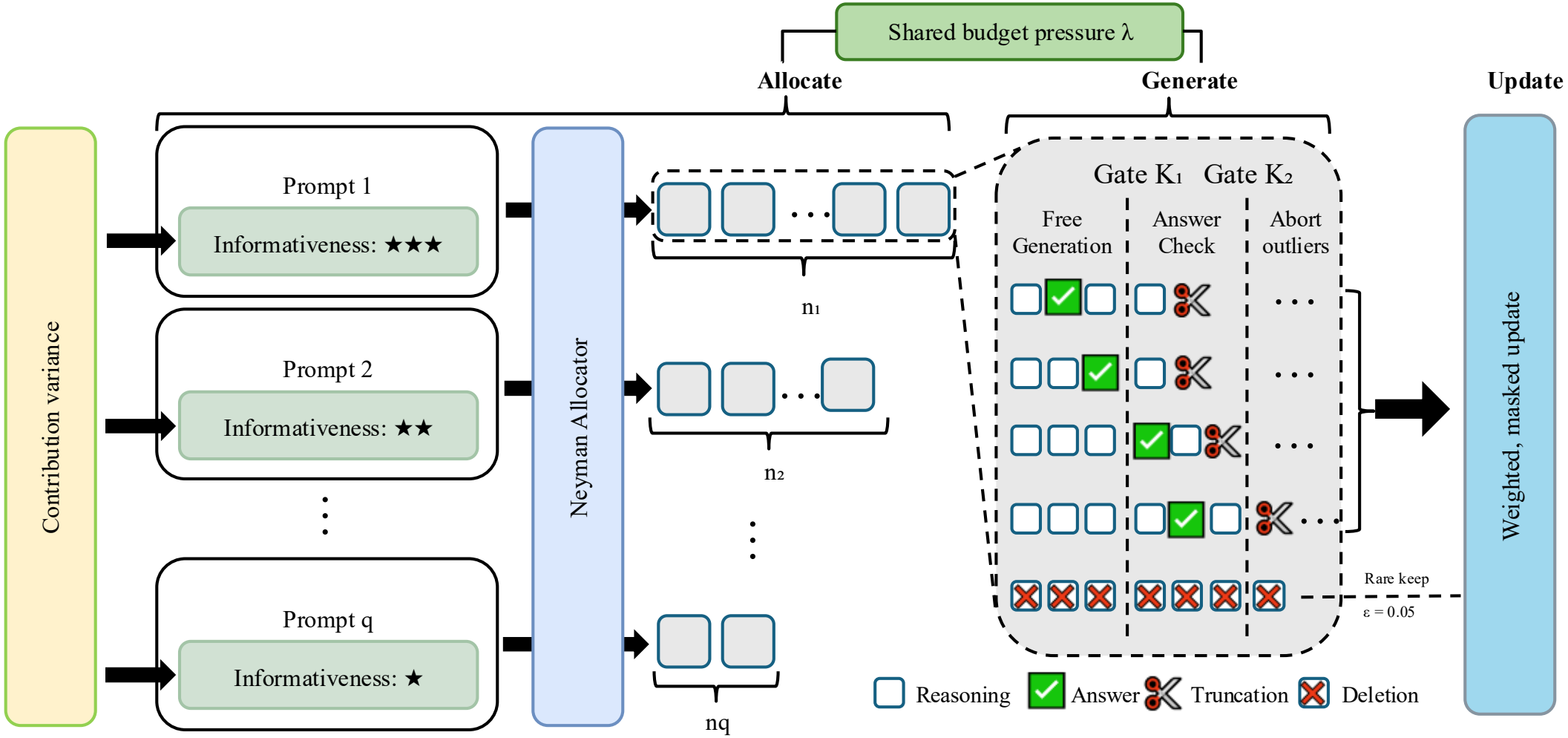
DUET: allocate → generate → update

2 Stop outliers that are unlikely to yield a meaningful answer



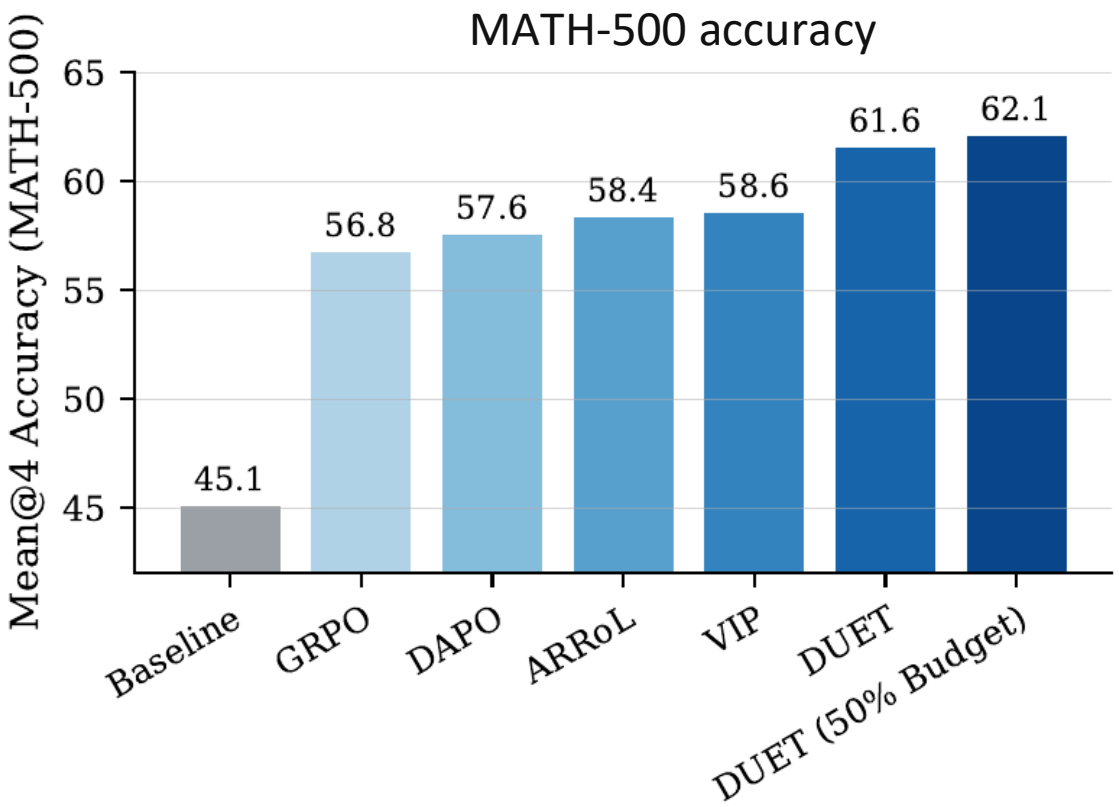
DUET: allocate → generate → update

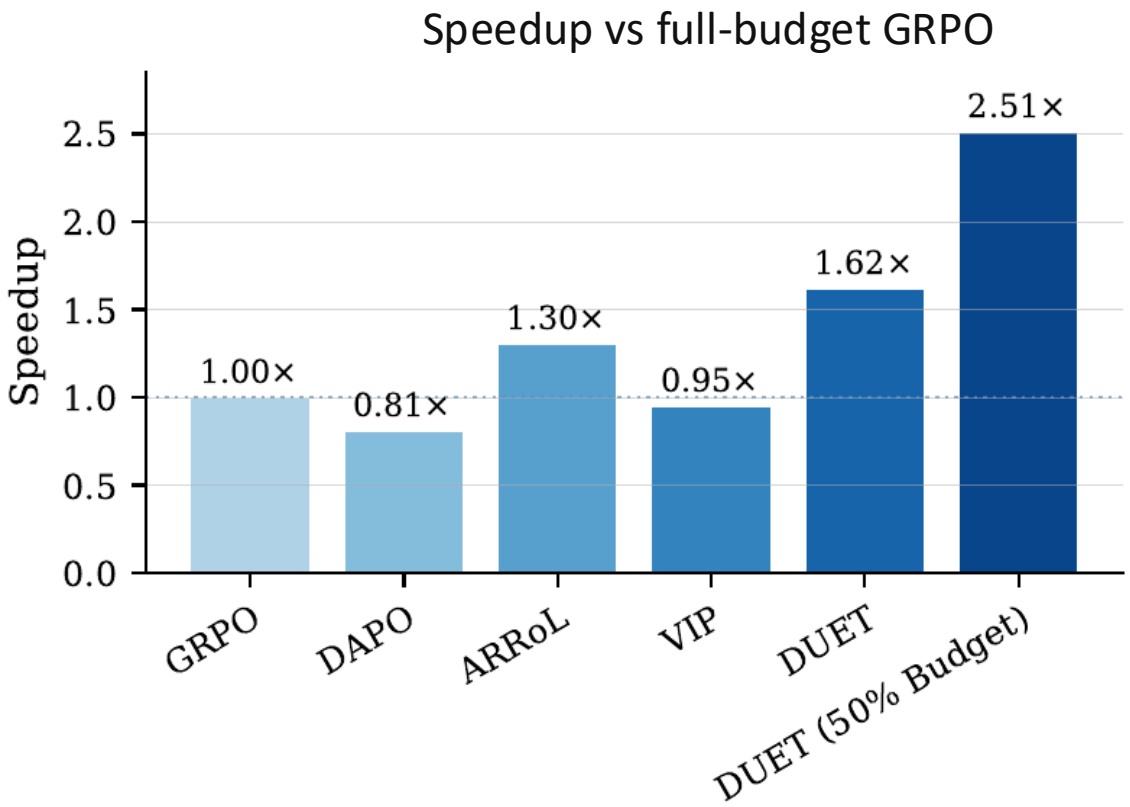
3 Shared budget, weighted updates; aborted traces contribute zero



Better accuracy with less training compute

Qwen3-1.7B · MATH training · same 8 × H100 setup

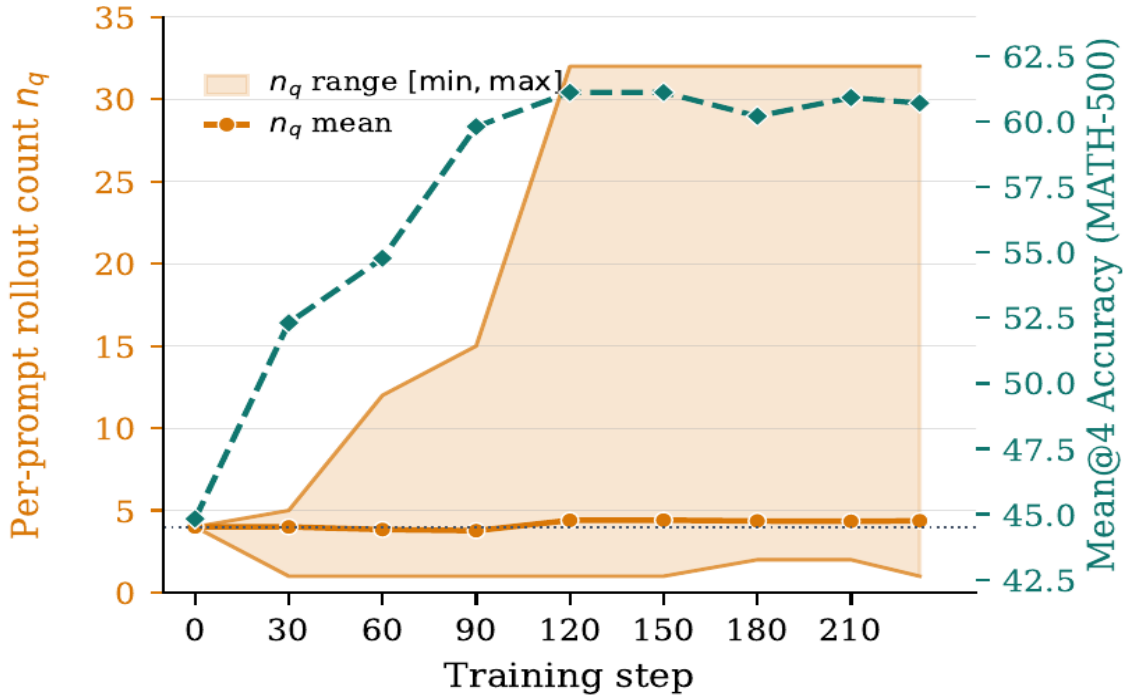




Half-budget DUET: +5.3 pp on MATH-500 and 2.51× speedup versus full-budget GRPO.

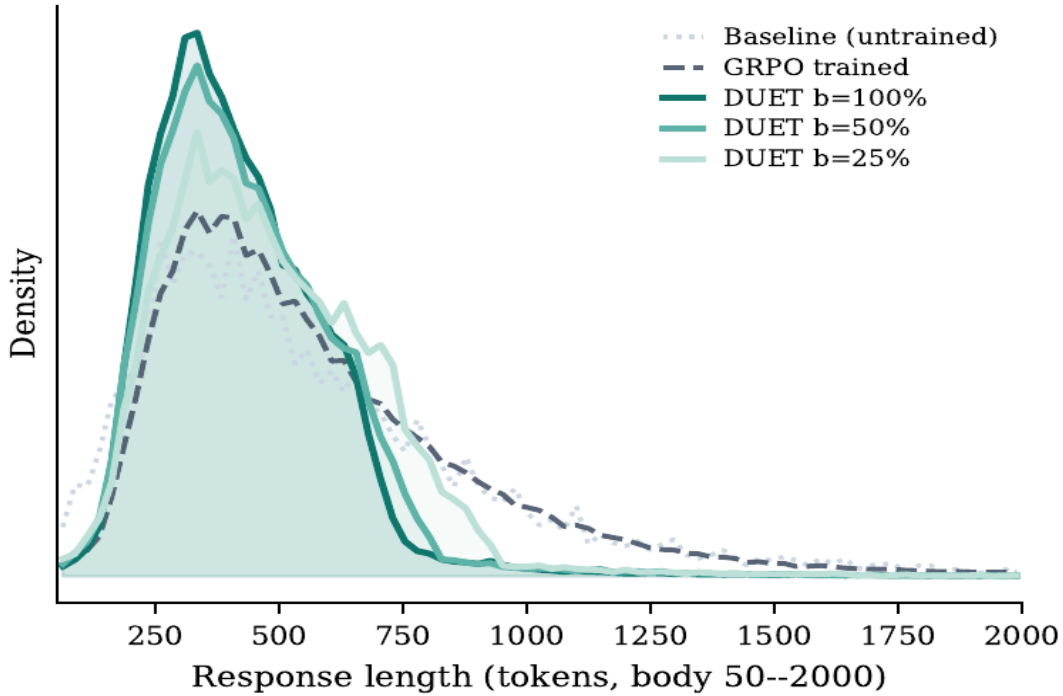
Emergent allocation and shorter reasoning

Allocation becomes heterogeneous



Uniform 4 rollouts → a learned range of 1–32

Responses become more concise



Median response: 511 → about 330 tokens
GRPO → DUET, alongside higher accuracy

Ablation: allocation and stopping play different roles

Qwen3-1.7B · all trained rows: 50% budget · speedup vs 50%-budget GRPO

Configuration	MATH-500	GSM8K	AIME-24	HumanEval	GPQA-D	Speedup
Untrained baseline	45.1	60.4	5.8	34.5	22.2	—
GRPO	56.4	75.8	6.7	46.9	27.4	1.00×
Allocation only	61.7	84.1	12.5	51.0	28.3	1.30×
Stopping only	61.3	82.1	12.5	49.4	28.6	1.93×
Both (DUET)	**62.1**	**84.0**	**10.8**	**51.7**	**29.1**	**1.73×**

Joint control reaches the highest MATH-500 accuracy.

Stopping alone is fastest; DUET balances quality and speed.

Robustness to controller choices

Changing the answer marker retains about 2× speedup

Final-checkpoint robustness runs · 50% budget · speed vs full-budget GRPO

Answer detector	Final MATH-500	Speedup
Boxed + answer phrases	**60.1 ± 0.5**	**2.09 ± 0.20×**
Strict boxed answer	60.4	1.91×
Language markers only	61.3	1.97×

Also tested: polling, prefix windows and cold-start settings.

Online vs static estimates: +2.87 pp at 25% budget

Compared with APRIL; combining asynchronous execution is future work.

DUET

Key takeaways

Coordinate allocation and stopping

Allocate by informativeness and expected cost. Stop stalled rollouts and reweight retained updates.

Improve the quality–efficiency trade-off

Achieve higher accuracy with less training compute in the evaluated RLVR settings.

Adapt as the policy learns

Learn heterogeneous rollout allocations and produce more concise responses.

More useful learning from a fixed budget

Paper

Code