SPEC-RL论文分享会

刘冰帅–厦门大学

2025.10.16

SPEC-RL: ACCELERATING ON-POLICY REINFORCEMENT LEARNING VIA SPECULATIVE ROLLOUTS

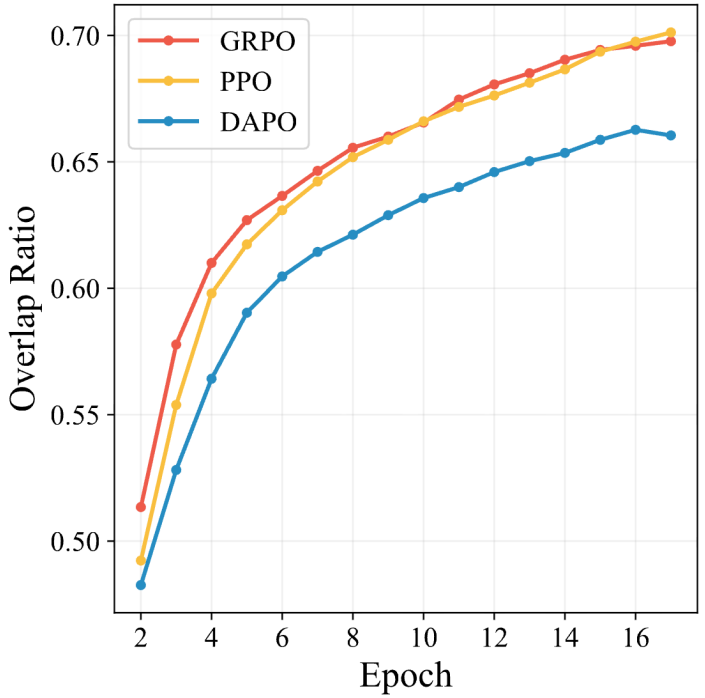
Bingshuai Liu[1,*], **Ante Wang**[1,3,*], **Zijun Min**[1,*], **Liang Yao**[2], **Haibo Zhang**[2], **Yang Liu**[3], **Anxiang Zeng**[2], **Jinsong Su**[1,†]

[1] School of Informatics, Xiamen University, [2] LLM Team, Shopee Pte. Ltd.,
[3] Institute for AI Industry Research (AIR), Tsinghua University

{bsliu,wangante,minzijun}@stu.xmu.edu.cn, {leon.yao,peter.wu}@shopee.com,
liuyang2011@tsinghua.edu.cn, zeng0118@ntu.edu.sg, jssu@xmu.edu.cn

Motivation

- 传统RLVR的问题
  - Rollout时间长，在不同setting下占据整体训练时间的50-70%
  - 因为zero-variance现象，一部分Rollout样本对于模型更新无效
- 此前的解决思路
  - DAPO: Dynamic Sampling
  - GRESO: Dynamic Sampling + 随机丢弃Rollout样本
  - 问题：没有真正地从Token粒度上加速Rollout
- 观察 & 思考
  - 相邻Epoch之间的Token overlap相当高 (最开始≈50%→最后≈70%)
  - 模型更新是一个渐进的过程，相邻epoch模型表现相当接近
  - 能否利用起来这种特性，加速Rollout?



Preliminary

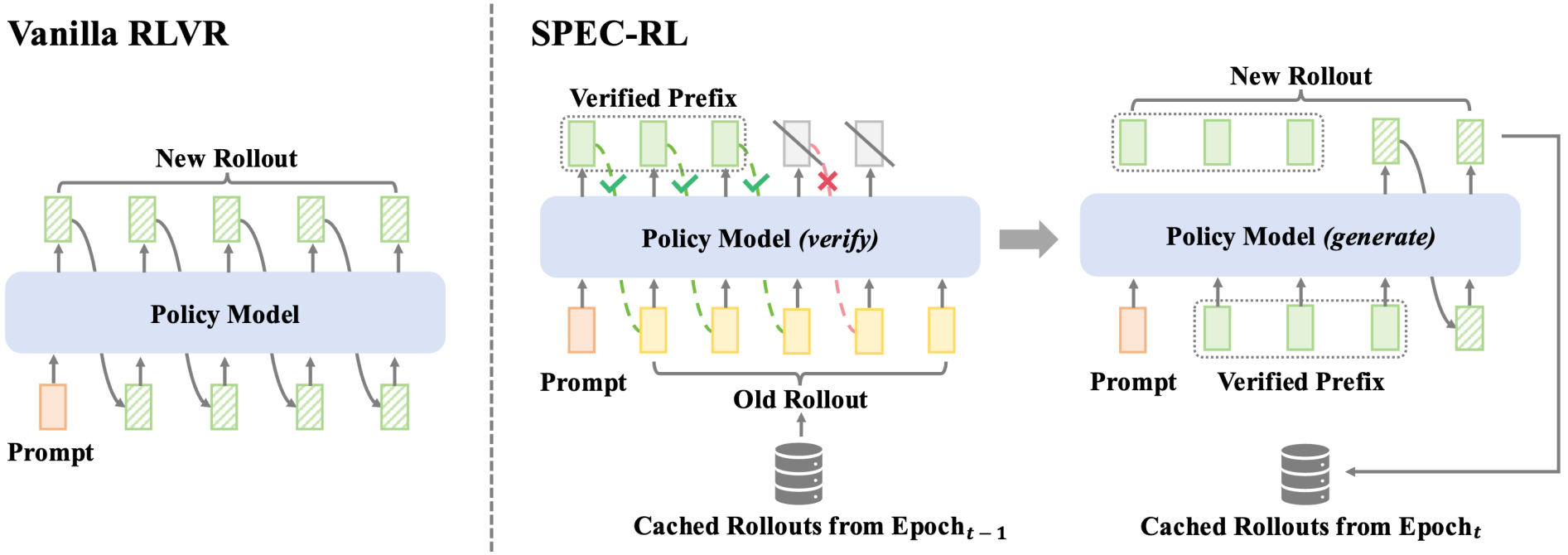
- 投机解码加速Inference (draft-and-verify)
  - Draft model: 草稿模型，比Target model更小，推理更快
  - Target model: 目标模型，用于并行verify草稿token
  - 多轮draft-and-verify: 比对两者对于相同token的概率，一旦拒绝，target model修正draft model，draft model重新继续生成草稿token，直到结束

```
[START] japan ' s benchmark bond n
[START] japan ' s benchmark nikkei 22 -5
[START] japan ' s benchmark nikkei 225 index rose 22 -6
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 - points
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 0 1
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 9859
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 - in
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in tokyo late
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in late morning trading . [END]
```

Fast Inference from Transformers via Speculative Decoding. ICML 2023 Oral

Method

- 将投机解码引入到RLVR
  - Old-policy作为draft model, 当前Policy作为target model
  - 直接读取old-policy的rollout结果和log-probs，无需引入额外的模型



Method

- 将投机解码引入到RLVR
  - 从第一个拒绝的位置开始，丢弃剩余token，让模型做继续生成
  - 引入宽容度系数ℓ，灵活控制当前policy对于old-policy的复用率
    - $\ell \to 1$，等价默认投机解码；$\ell \to \infty$, 等价完全复用old-policy

Algorithm 1: SPEC-RL

Input: Current policy π_t; Prompt $\mathbf{x}$; old response $\mathbf{y}^{old} = \{y_i^{old}\}$ with probability p^{old}; lenience $\ell \geq 1$.

Compute probability *in parallel* $p_i^{new} \leftarrow \pi_t(y_i^{old} \mid \mathbf{x}, \mathbf{y}_{<i}^{old}), i = 1, ..., |\mathbf{y}^{old}|$;

Compute acceptance probability $\tilde{\alpha} = \min(1, \ \ell \cdot \frac{p^{new}}{p^{old}})$;

Initialize rejection position $n \leftarrow |\mathbf{y}^{old}| + 1$;

for $i = 1$ **to** $|\mathbf{y}^{old}|$ **do**

  Sample $u \sim \mathcal{U}(0, 1)$;

  if $u > \tilde{\alpha}_i$ **then**

    Assign rejection position $n \leftarrow i$;

    break;

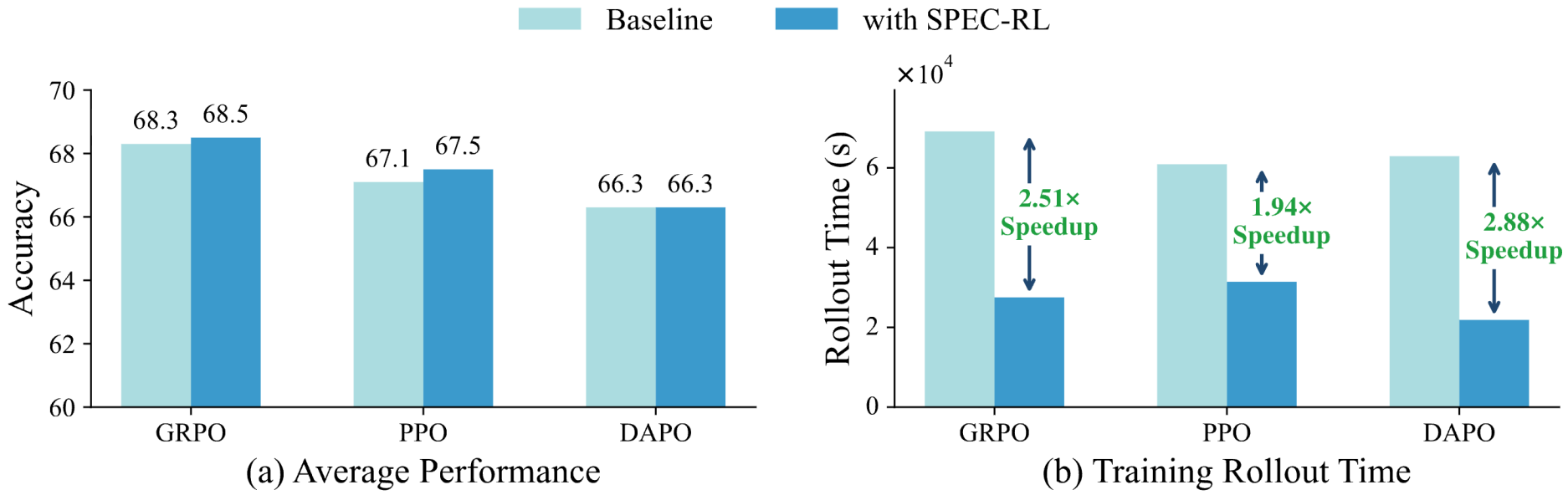
Generate response $\mathbf{y}_{\geq n}^{new} \leftarrow \pi_t(\cdot \mid \mathbf{x}, \mathbf{y}_{<n}^{old})$;

Assemble response $\mathbf{y}^{new} \leftarrow \{\mathbf{y}_{<n}^{old}, \mathbf{y}_{\geq n}^{new}\}$;

return $\mathbf{y}^{new}$

Experiments

- 整体性能
  - 配置：Qwen-3-8B-Base, GRPO/PPO/DAPO
  - 在维持原本推理能力的基础上，实现2-3倍Rollout加速



Experiments

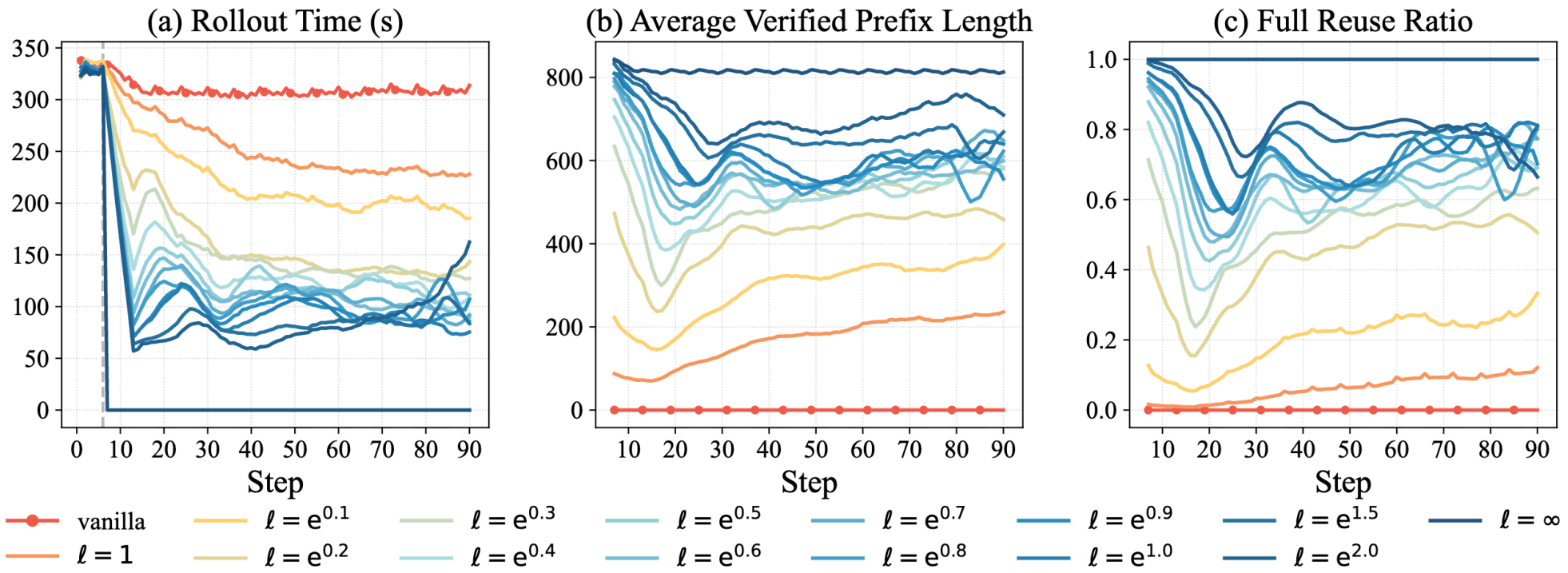
整体性能

配置：Qwen-3-1.7B/8B-Base, LLaMA-3.2-1B-Instruct, GRPO/PPO/DAPO

Algorithm	Rollout Efficiency		Math Reasoning					OOD		AVG
	Tokens (M)	Speedup	AMC23	GSM8K	MATH 500	Minerva Math	Olympiad Bench	MMLU STEM	IFEval	
Qwen-3-1.7B-Base										
Base Model	-	-	22.5	59.1	45.0	12.5	16.7	39.3	17.9	30.4
GRPO	554.8	1.00×	**42.5**	82.6	64.4	26.5	25.5	**60.7**	24.4	46.7
↪ + SPEC-RL	**182.7**	**2.29**×	37.5	**84.4**	**68.0**	**29.4**	**29.3**	58.3	**28.8**	**48.0**
PPO	565.1	1.00×	**35.0**	**82.0**	63.0	**26.8**	25.3	**59.4**	25.5	45.3
↪ + SPEC-RL	**230.8**	**1.94**×	**35.0**	**82.0**	**64.8**	25.4	**25.9**	58.6	**25.9**	**45.4**
DAPO	543.1	1.00×	**30.0**	79.6	**60.8**	24.6	23.0	52.2	24.8	**42.1**
↪ + SPEC-RL	**171.6**	**2.17**×	22.5	**80.1**	60.0	**25.7**	**25.5**	**53.5**	**27.0**	42.0
Qwen-3-8B-Base										
Base Model	-	-	40.0	83.0	67.4	27.2	34.1	60.4	29.9	48.9
GRPO	1033.1	1.00×	**75.0**	94.1	86.4	43.8	**53.0**	**84.6**	41.2	68.3
↪ + SPEC-RL	**336.6**	**2.51**×	70.0	**94.5**	**87.8**	**44.1**	51.0	84.5	**47.7**	**68.5**
PPO	984.0	1.00×	70.0	**94.2**	**85.8**	43.0	**51.6**	83.8	**41.6**	67.1
↪ + SPEC-RL	**400.1**	**1.94**×	**75.0**	92.9	85.2	**43.4**	50.8	**84.4**	41.0	**67.5**
DAPO	1052.2	1.00×	**75.0**	93.3	**84.8**	40.1	48.6	**82.4**	39.6	**66.3**
↪ + SPEC-RL	**326.2**	**2.88**×	65.0	**93.8**	84.4	**43.8**	**50.4**	82.2	**44.4**	**66.3**
LLaMA-3.2-1B-Instruct										
Base Model	-	-	0.0	26.7	14.2	4.0	2.8	32.6	37.0	16.8
GRPO	553.9	1.00×	5.0	28.1	19.2	**3.3**	4.9	33.1	37.0	18.7
↪ + SPEC-RL	**162.5**	**2.60**×	**7.5**	**28.7**	**19.4**	1.8	**5.0**	**34.5**	**37.2**	**19.2**
PPO	521.5	1.00×	**10.0**	31.6	**20.8**	4.0	**6.4**	34.3	**42.7**	**21.4**
↪ + SPEC-RL	**210.6**	**2.01**×	**10.0**	**32.4**	20.2	**5.5**	5.0	**35.3**	40.7	21.3
DAPO	482.6	1.00×	7.5	29.6	19.2	**4.0**	**5.5**	33.0	**38.6**	19.6
↪ + SPEC-RL	**123.1**	**2.48**×	**10.0**	**34.9**	**20.2**	**4.0**	**5.5**	**35.5**	38.4	**21.2**

Experiments

- 关于宽容度ℓ的消融实验
  - 配置：Qwen-3-1.7B-Base, 宽容度$\ell = 1, e^{0.1}, e^{0.2}, \ldots, \infty$，GRPO
  - Rollout时间随着宽容度逐渐提升明显下降、平均复用token数量也越来越多



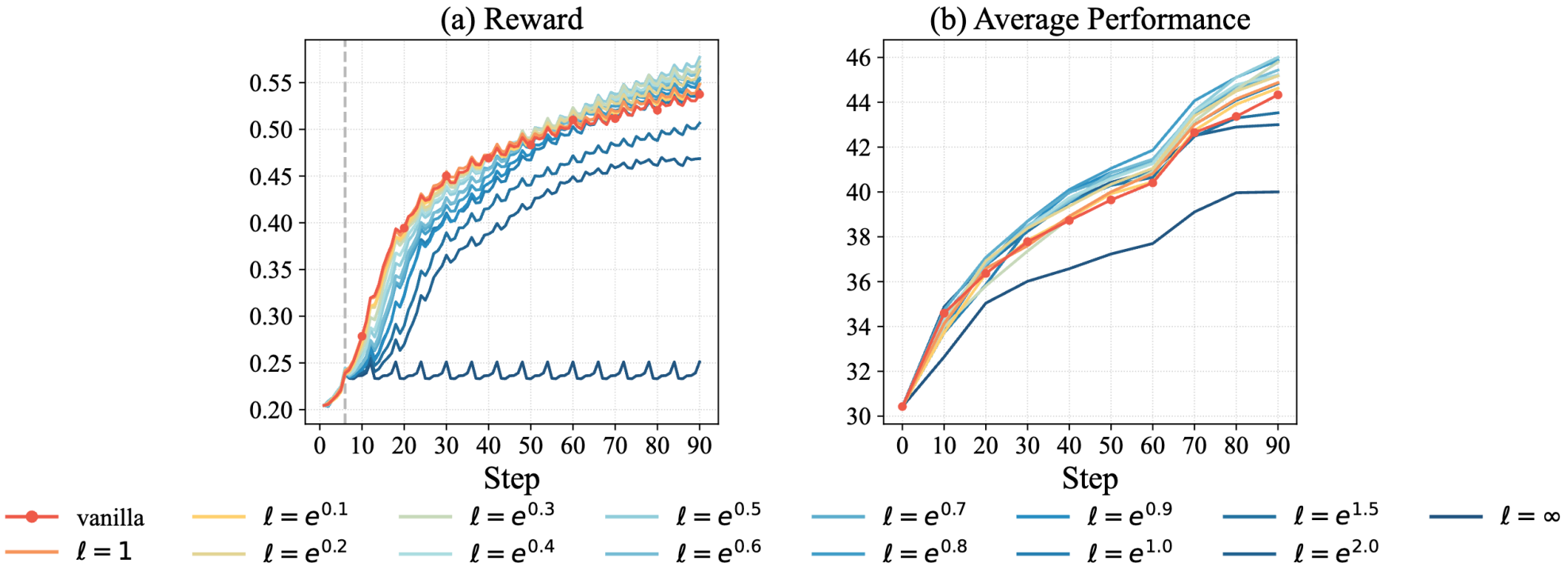
Experiments

- 关于宽容度ℓ的消融实验
  - 合适的宽容度能够很好地平衡加速比和学习效果 ($\ell = e^{0.5}$)
  - 过度复用尽管能达到相当高的加速比，但是会导致模型彻底崩溃 ($\ell = e^{2.0}, \infty$)

Algorithm	Rollout Efficiency		Math Reasoning					OOD		AVG
	Tokens (M)	Speedup	AMC23	GSM8K	MATH 500	Minerva Math	Olympiad Bench	MMLU STEM	IFEval	
GRPO	554.8	1.00×	**42.5**	82.6	64.4	26.5	25.5	60.7	24.4	46.7
↪ + SPEC-RL $\ell = 1$	419.1	1.22×	40.0	81.8	63.8	28.7	26.5	59.6	25.9	46.6
↪ + SPEC-RL $\ell = e^{0.2}$	246.7	1.86×	37.5	83.3	66.4	**29.8**	**29.6**	58.5	25.9	47.3
↪ + SPEC-RL $\ell = e^{0.5}$	182.7	2.29×	37.5	**84.4**	**68.0**	29.4	29.3	58.3	28.8	**48.0**
↪ + SPEC-RL $\ell = e^{0.8}$	144.8	2.64×	37.5	83.5	63.6	27.2	25.0	**61.7**	26.2	46.4
↪ + SPEC-RL $\ell = e^{1.0}$	123.0	2.91×	37.5	83.9	62.4	25.7	24.9	54.8	28.3	45.4
↪ + SPEC-RL $\ell = e^{2.0}$	114.4	3.05×	30.0	80.4	55.0	21.0	21.9	53.5	**29.0**	41.5
↪ + SPEC-RL $\ell = \infty$	**40.0**	**14.86**×	32.5	78.1	60.4	19.9	23.7	44.1	22.0	40.1

Experiments

- 关于宽容度ℓ的消融实验
  - 合适的宽容度能够很好地平衡加速比和学习效果 ($\ell = e^{0.5}$)
  - 合适的宽容度的Reward和平均推理性能可以在训练中后期超越baseline算法



Experiments

端到端的时间分析

- SPEC-RL在只修改Rollout过程的基础上，最佳情况实现端到端时间减半
- DAPO + SPEC-RL: 从24.2h →12.9h

Algorithm	End-to-end (h)	Average step time (s)												
	Total	Total	Δ vs. base	verification	rollout	assembly	reward	old-log-probs	ref	values	adv	update-critic	update-actor	others
Qwen-3-1.7B-Base														
GRPO	12.63	505.1	–	–	309.9	–	91.0	17.2	15.8	–	0.4	–	56.0	14.9
↪ + SPEC-RL	8.65	346.0	↓ 159.1	22.1	135.2 (2.29×)	1.5	81.0	17.1	16.3	–	0.5	–	56.2	16.2
PPO	14.10	563.9	–	–	308.1	–	100.5	17.2	–	14.0	4.7	46.0	56.5	16.9
↪ + SPEC-RL	10.78	431.2	↓ 132.7	22.7	158.6 (1.94×)	1.4	94.1	17.3	–	13.8	4.6	45.0	55.5	18.1
DAPO	11.10	443.8	–	–	301.3	–	93.1	8.6	–	–	0.3	–	25.9	14.6
↪ + SPEC-RL	7.90	316.0	↓ 127.9	21.0	139.0 (2.17×)	1.4	97.9	18.1	–	–	0.2	–	25.9	12.7
Qwen-3-8B-Base														
GRPO	31.66	1266.4	–	–	768.2	–	73.2	66.8	66.9	–	4.2	–	263.8	23.4
↪ + SPEC-RL	21.03	841.0	↓ 425.4	74.7	305.8 (2.51×)	1.3	61.4	63.8	62.4	–	4.9	–	248.8	18.0
PPO	34.85	1393.9	–	–	676.7	–	70.5	65.4	–	57.4	4.2	224.1	260.4	35.3
↪ + SPEC-RL	26.97	1078.8	↓ 315.1	71.5	349.3 (1.94×)	1.4	64.9	59.6	–	52.1	4.9	205.9	236.9	32.5
DAPO	24.29	971.8	–	–	699.2	–	64.4	66.3	–	–	0.1	–	121.1	20.7
↪ + SPEC-RL	12.90	515.9	↓ 455.9	51.0	243.0 (2.88×)	1.1	54.0	51.2	–	–	0.1	–	97.5	18.0
LLaMA-3.2-1B-Instruct														
GRPO	10.20	408.0	–	–	229.7	–	105.8	12.6	11.5	–	0.4	–	34.7	13.2
↪ + SPEC-RL	7.28	291.3	↓ 116.7	17.2	88.3 (2.60×)	1.4	110.4	13.0	11.9	–	0.5	–	34.4	14.4
PPO	10.94	437.6	–	–	218.9	–	117.6	12.5	–	10.0	4.8	10.0	32.6	31.3
↪ + SPEC-RL	8.60	344.0	↓ 93.6	17.5	108.9 (2.01×)	1.3	110.9	12.4	–	10.1	4.6	10.1	34.3	33.8
DAPO	9.77	328.4	–	–	198.4	–	100.8	11.2	–	–	0.1	–	9.6	8.5
↪ + SPEC-RL	6.97	238.4	↓ 90.0	13.4	80.0 (2.48×)	1.1	110.5	11.5	–	–	0.1	–	9.9	12.0

Case Study

Prompt

system
You are a helpful assistant.
user
If $-3 \leq x + 5 \leq 8$ and x is an integer, what is the sum of all the possible solutions?
Please reason step by step, and put your final answer within \boxed{}.
assistant

Old Rollout from Epoch$_{t-1}$

To solve the given inequality $-3 \leq x + 5 \leq 8$, we need to isolate x on one side of the inequality. We can do this by subtracting 5 from all parts of the inequality:

$$-3 - 5 \leq x + 5 - 5 \leq 8 - 5$$

Simplifying each part:

$$-8 \leq x \leq 3$$

Since x is an integer, the possible values for x are:

$$\{-8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3\}$$

Next, we sum these possible values:

$$-8 + (-7) + (-6) + (-5) + (-4) + (-3) + (-2) + (-1) + 0 + 1 + 2 + 3$$

We can group the positive and negative terms:

$$(-8 + 3) + (-7 + 2) + (-6 + 1) + (-5 + 4) - 10 = -5 - 5 - 5 - 10 = -25$$

Thus, the sum of all the possible solutions is:
\boxed{-25}.

New Rollout from Epoch$_{t}$

To solve the given inequality $-3 \leq x + 5 \leq 8$, we need to isolate x on one side of the inequality. We can do this by subtracting 5 from all parts of the inequality:

$$-3 - 5 \leq x + 5 - 5 \leq 8 - 5$$

Simplifying each part:

$$-8 \leq x \leq 3$$

Since x is an integer, the possible values for x are:

$$\{-8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3\}$$

Next, we need to find the sum of all these possible values of x. We can sum them directly:

$$-8 + (-7) + (-6) + (-5) + (-4) + (-3) + (-2) + (-1) + 0 + 1 + 2 + 3$$

We can simplify this by grouping the terms to make the addition easier:

$$(-8 + 3) + (-7 + 2) + (-6 + 1) + (-5 + 0) + (-4 + (-2)) + (-3 + (-1)) = -5 + (-5) + (-5) + (-5) + (-6) + (-4) = -30$$

Thus, the sum of all the possible solutions is:
\boxed{-30}.

Discussion & Further Work

- SPEC-RL和传统算法的区别
  - SPEC-RL & 投机解码: 传统投机解码涉及到多轮，并且需要显示引入一个draft model，SPEC-RL只需要单轮，并且直接读取old-policy的rollout结果，无需额外引入draft model，保证与RLVR算法其他部分解耦
  - SPEC-RL & RLVR: 传统RLVR每次都需要从头重新Rollout，时间上成为整体训练pipeline上的瓶颈，SPEC-RL通过复用old-policy，采用宽容度ℓ控制复用程度，在大幅削减Rollout时间的同时保证整体学习效果
- 未来方向
  - SPEC-RL只加速Rollout过程，并且设计上和RLVR的其他流程完全解耦，与其他方向的RLVR加速工作是正交的，未来将其与SPEC-RL结合起来可以进一步削减端到端时间

📁 **Code:** https://github.com/ShopeeLLM/Spec-RL

谢 谢!

Contact: bsliu@stu.xmu.edu.cn
jssu@xmu.edu.cn