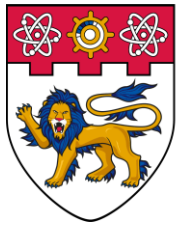


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

ReSpec: Towards Optimizing Speculative Decoding in Reinforcement Learning Systems

Qiaoling Chen, Zijun Liu , Peng Sun , Shenggui Li , Guoteng Wang , Ziming Liu , Yonggang Wen , Siyuan Feng , Tianwen Zhang

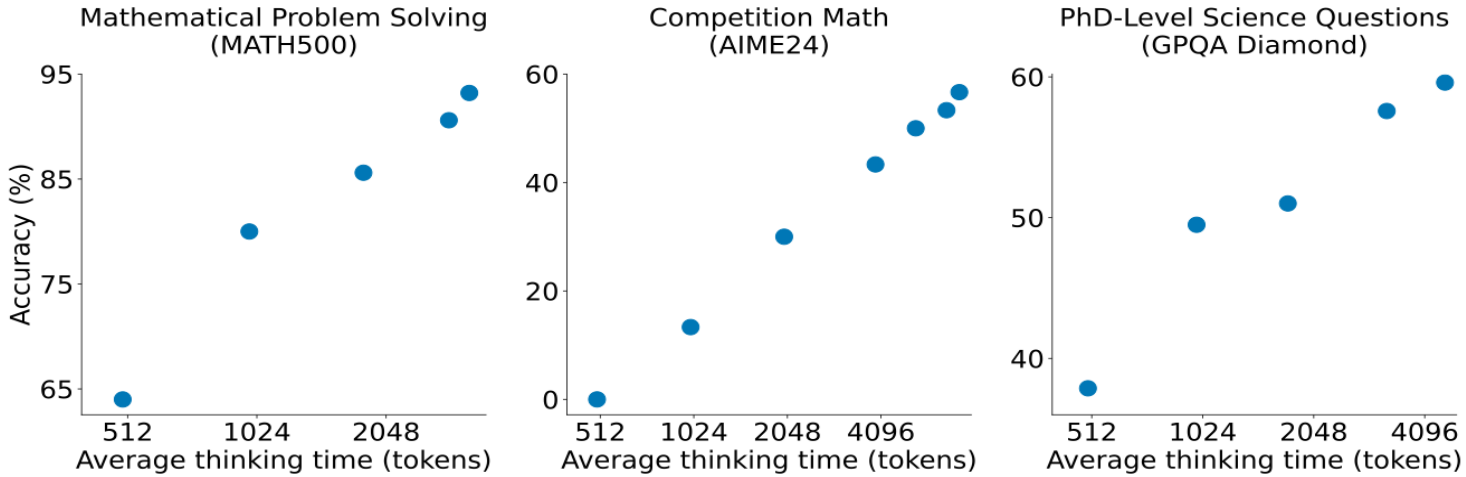
Nanyang Technological University , Shanghai Qiji Zhifeng Co., Ltd.

The Rise of Reasoning LLMs

- Reasoning LLMs show strong capability in math, coding. Etc

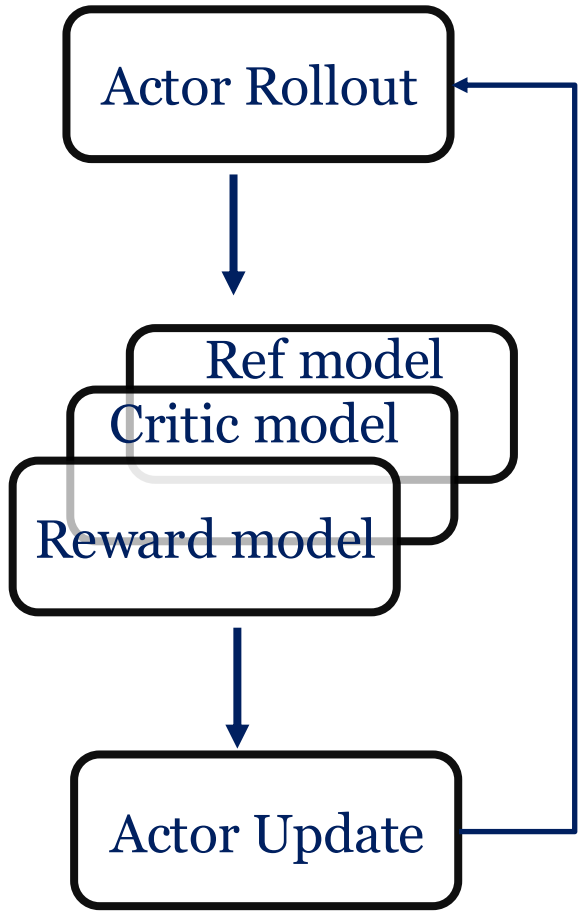
Claude Gemini Qwen

- Longer response better performance



Advancing Reasoning LLMs via RL

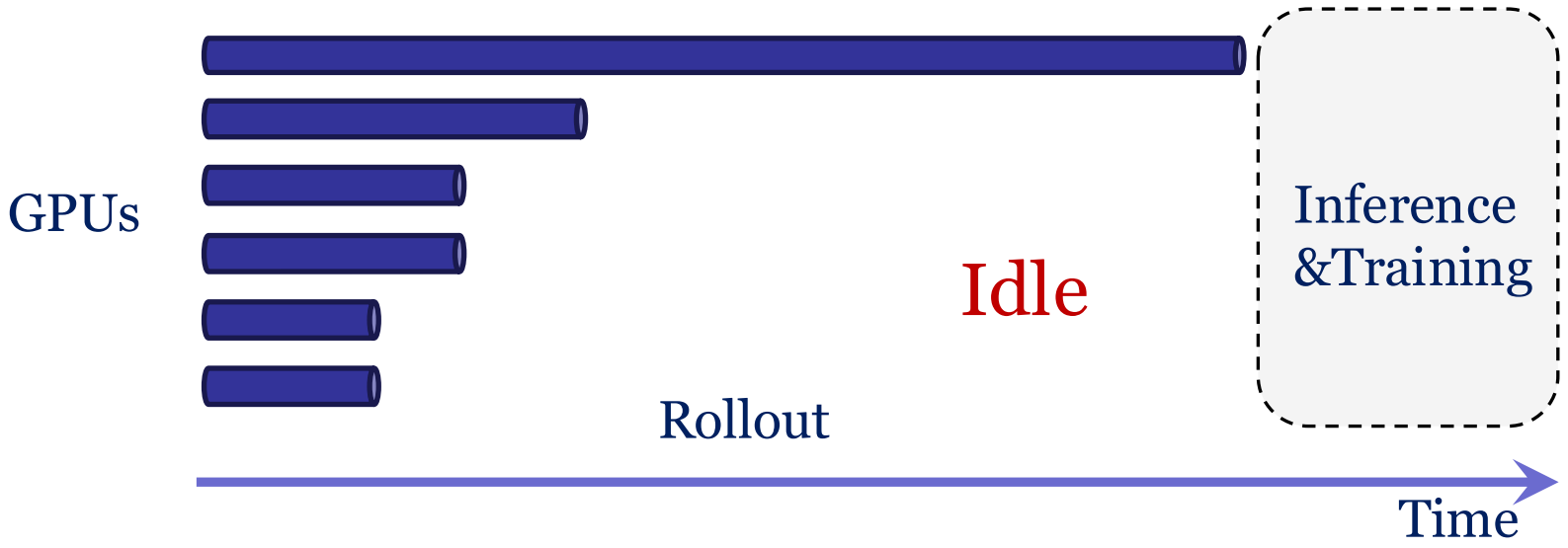
- Step 1. Rollout
  - ➢Generate multiple response for each prompts
- Step 2. Inference, rule-base reward
  - ➢Assign reward score based on reward function
- Step 3. Model training, policy optimization



Long tail rollout in RL

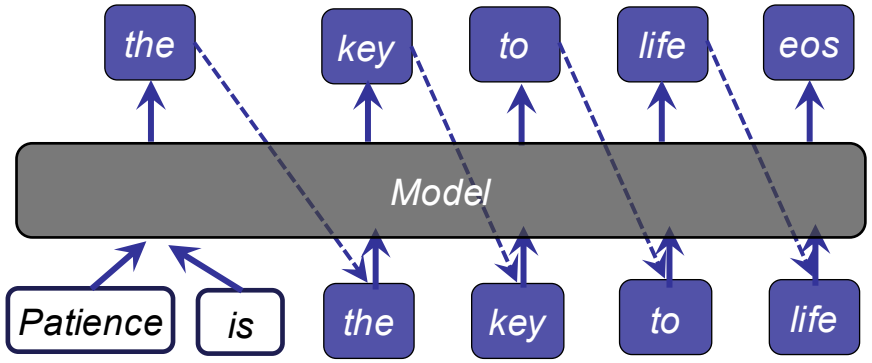
- Rollout stage dominate end-to-end RL iteration time

	Generation	Inference	Training
Math	83%–86%	1.7%–2.4%	12.3%–14.6%
Code	70.9%–75.5%	11.4%–14%	13.1%–15.1%



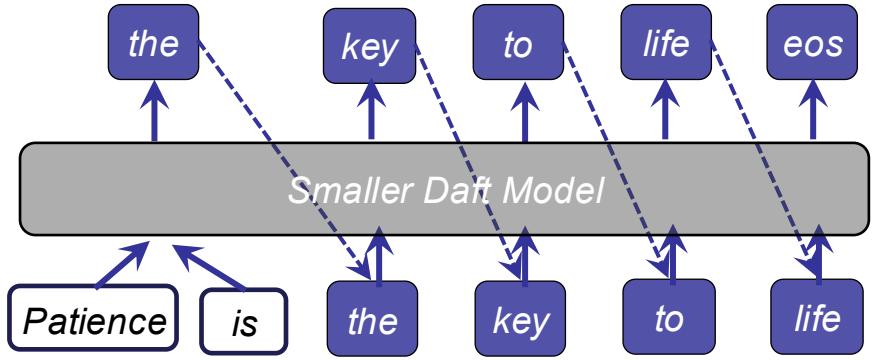
Speculative Decoding for generation

Vanilla decoding

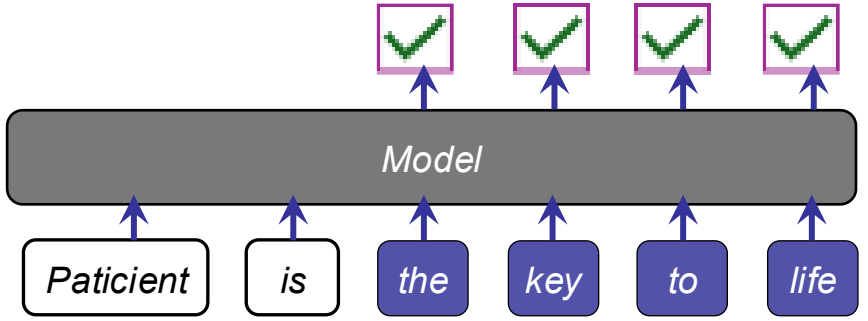


Speculative decoding

Step1: Drafting

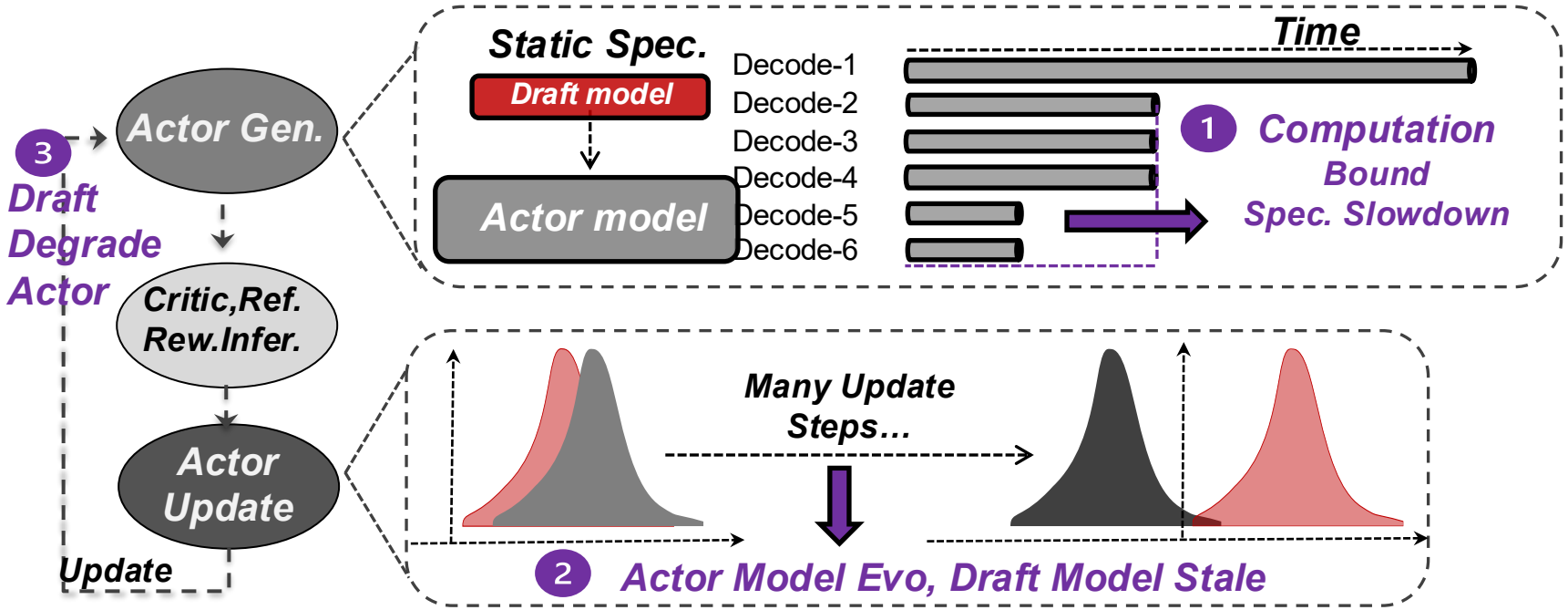


Step2: Verification



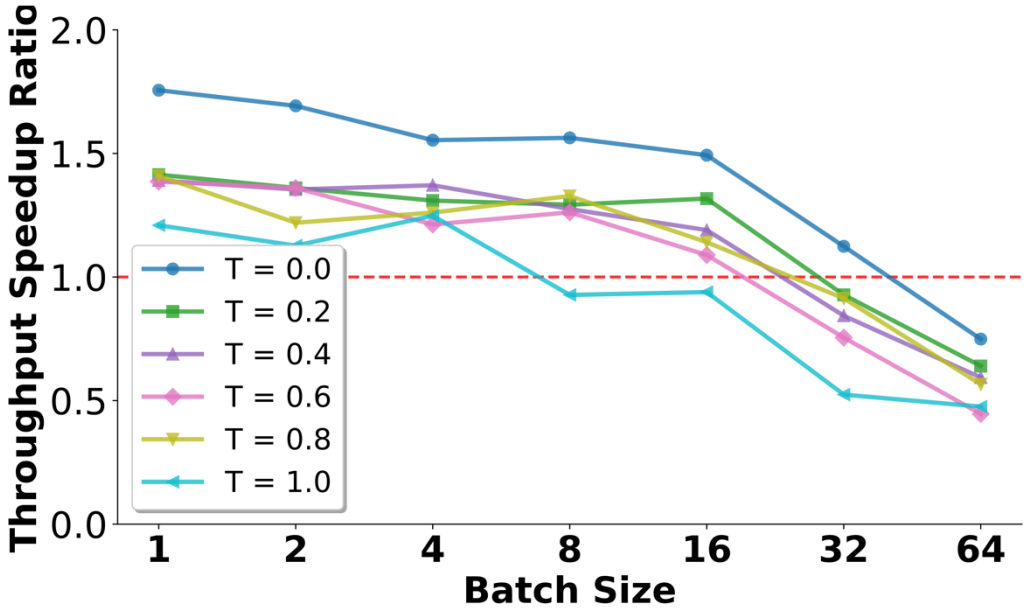
Challenges of Applying SD in RL

- GAP1: Diminishing speedup at large batch sizes.
- GAP2: SD staleness during RL training.
- GAP3: Drafter-induced degradation of the actor's performance.

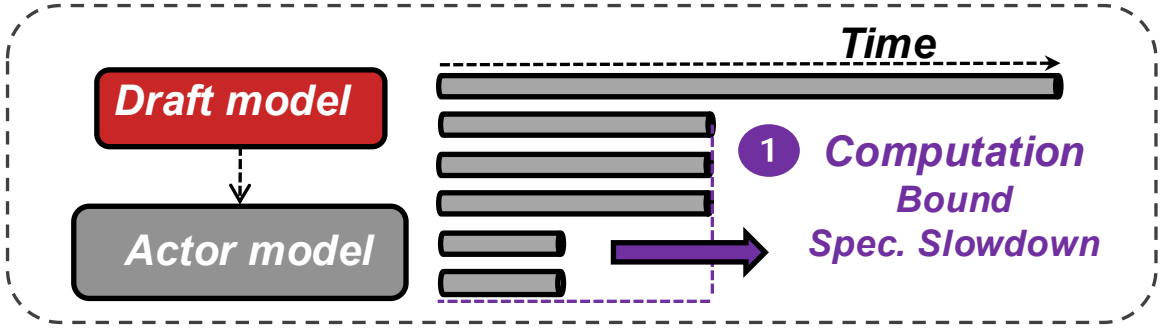


Challenges of Applying SD in RL

- GAP1: Diminishing speedup at large batch sizes.



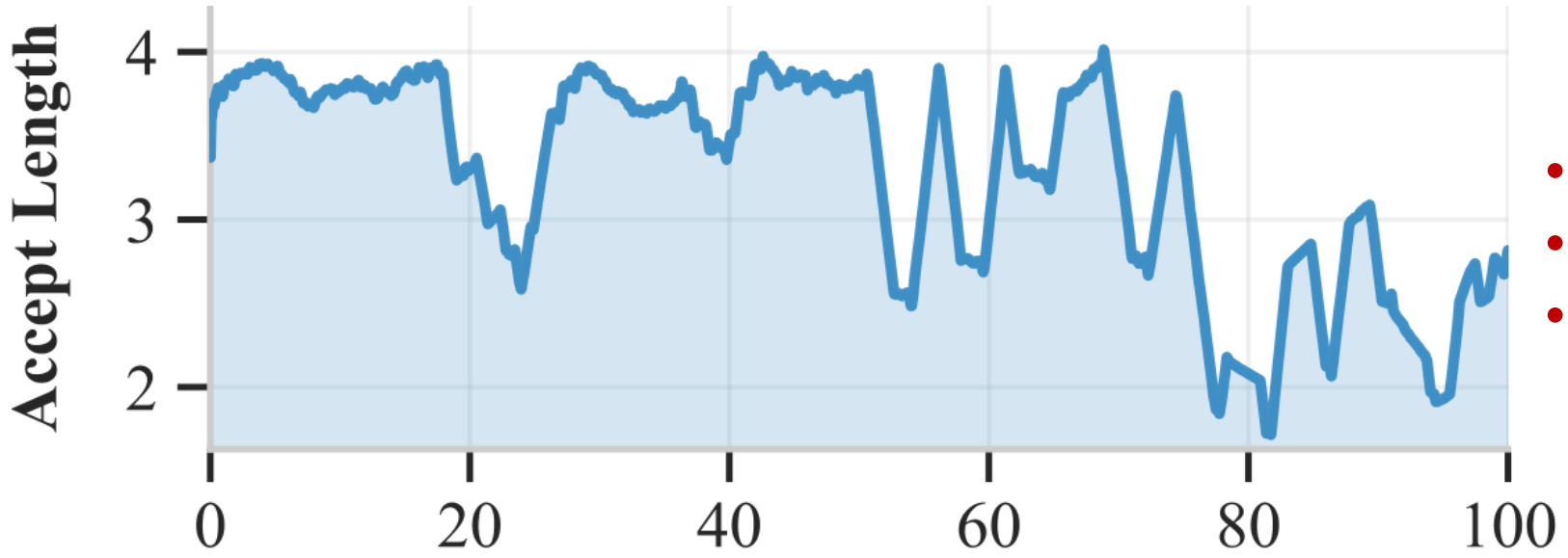
- *Qwen2.5-7B Instruct*
- *MT-bench*



High Verification Cost

Challenges of Applying SD in RL

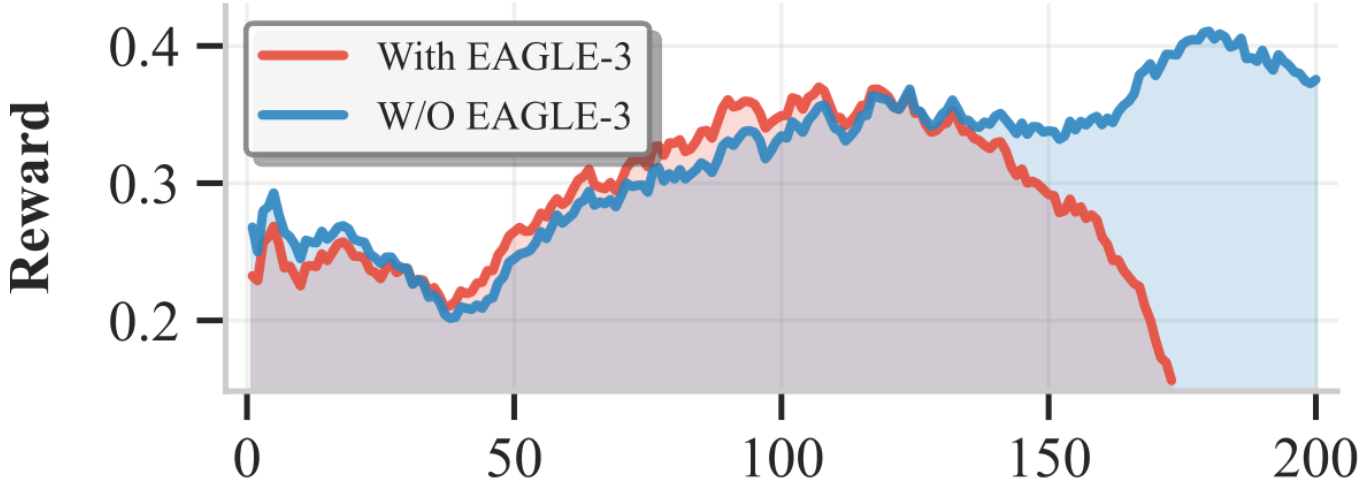
- GAP2: SD staleness during RL training.



- *Qwen2.5-7B-Instruct*
- *Math*
- *RL training*

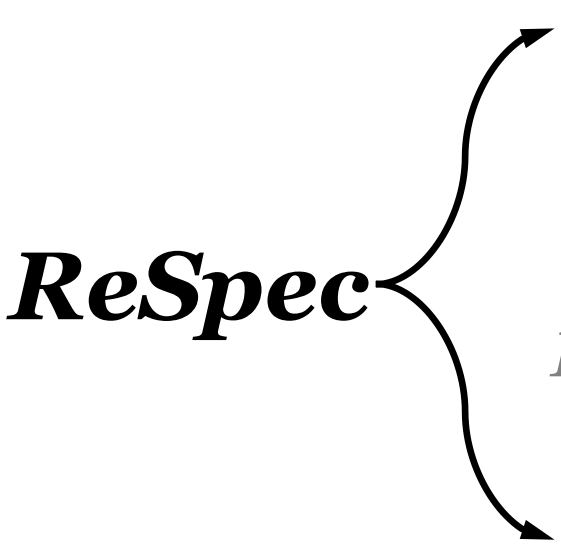
Challenges of Applying SD in RL

- GAP3: Drafter-induced degradation of the actor's performance.



- R1: non-deterministic verification paths
- R2: drafter staleness (compounding Gap 2)
- R3: variance amplification under non-stationary updates

ReSpec Design



Rollout ***Adaptive Speculative Decoding***

Dynamically switches between speculative and non-speculative execution modes according to the current system state.

Inference

Training ***Online Learner***

Maintains the draft model through online training by on-policy signals and knowledge distillation.

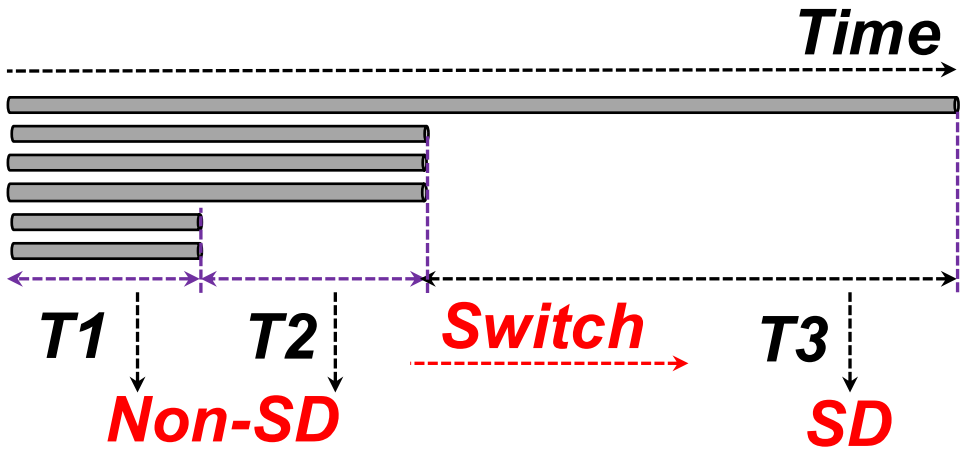
Adaptive Speculative Decoding

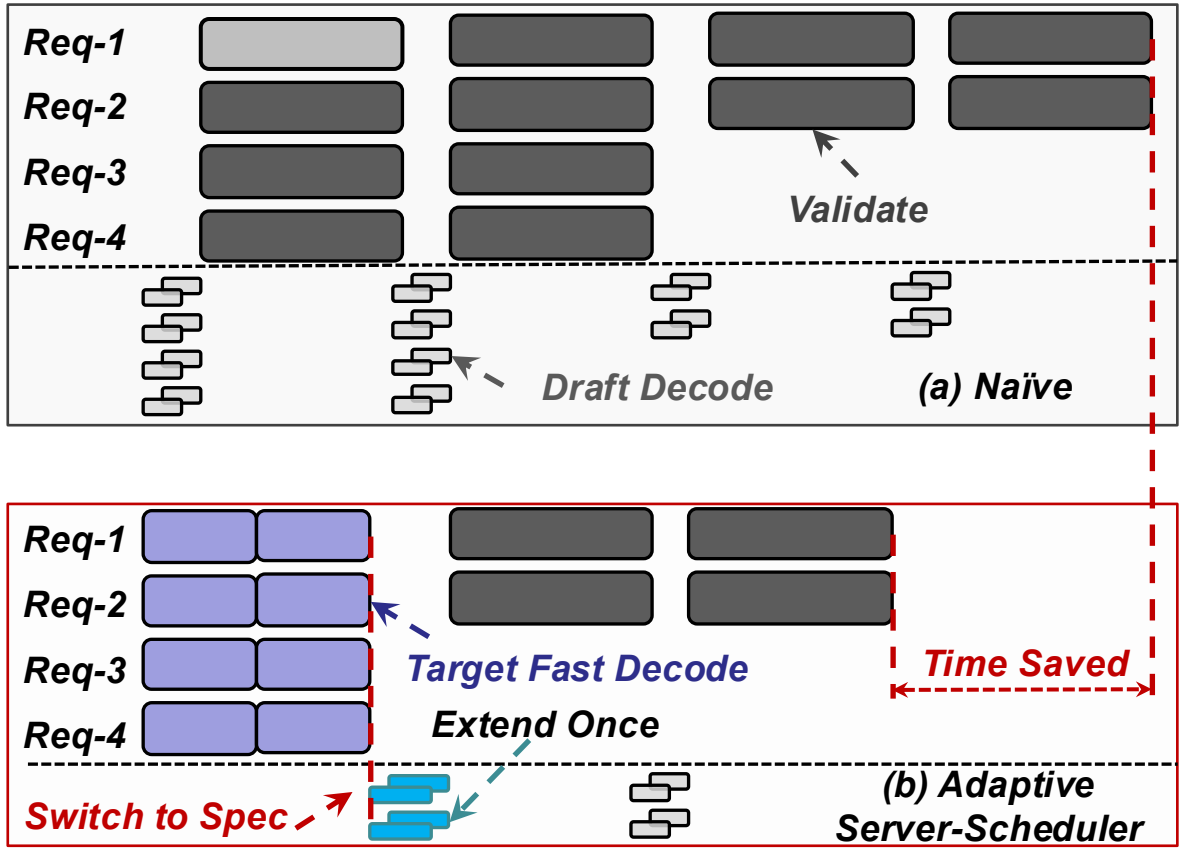
① Solver:

Estimates the throughput speedup of different SD configurations as a function of the active batch size.

② Scheduler:

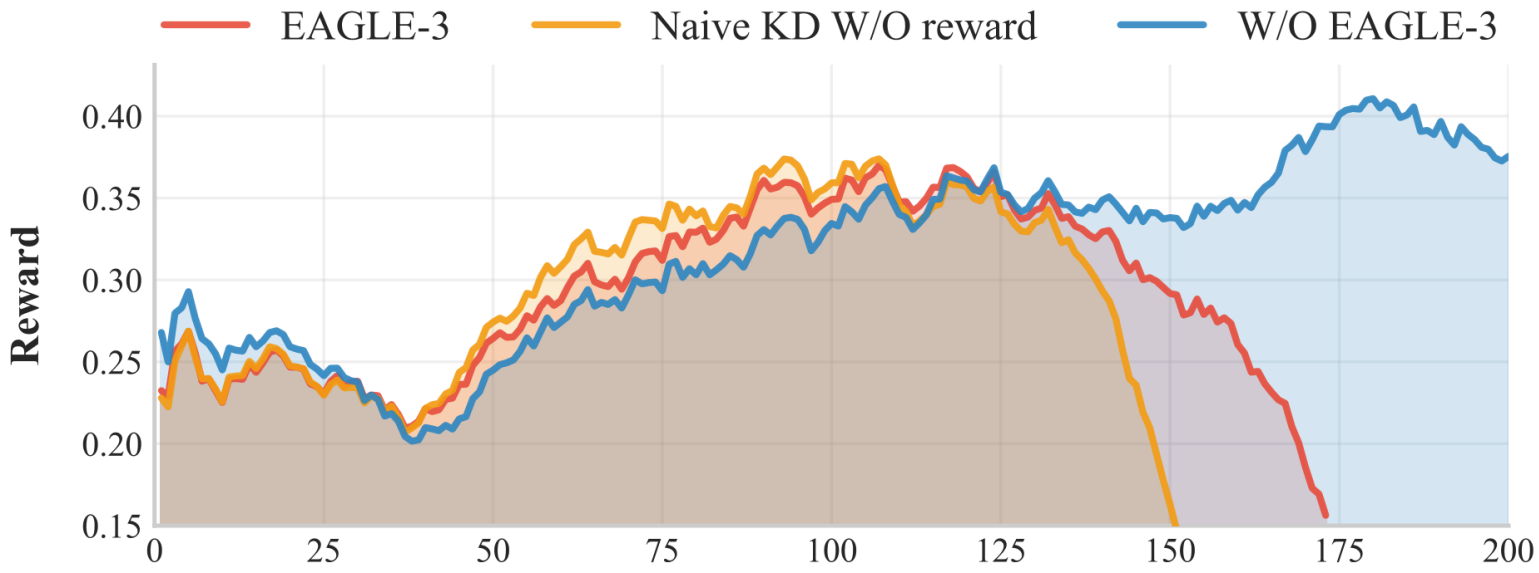
Switching between speculative and non-speculative execution without extra overhead.





Online Learner: naïve KD

- Naïve KL loss between Target's & Draft's logits
  - ➢Collapse quicker
    - →No-reward signal
    - →Draft model update wrong



Online Learner: reward-weighted KD

- Reward-weighted KD Algorithmn for updating Draft model
  - Variable from rollout
    - X -> input
    - Y -> output
    - q -> Target model logist
    - p -> Draft model logist
    - w(r) -> Reward Score from inference

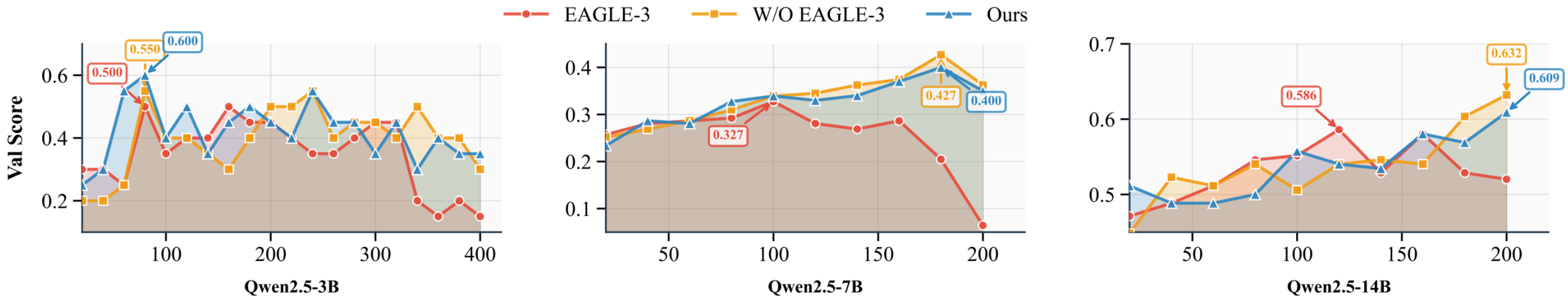
$$\mathcal{L}_{\mathcal{KD}}(x, y) = w(r) \sum_{t=1}^{T} \mathrm{KL}(\tilde{p}(\cdot \mid x, y_{<t}) \,\|\, q_\theta(\cdot \mid x, y_{<t}))$$

Evaluation

- Model:
  - Qwen 2.5 - ***3B, 7B,14B***
- Algorithmn:
  - GRPO for RL
  - EAGLE-3 for SD
- Training and Inference Engine
  - Verl & SGLang
- Infrastrature
  - 64× NVIDIA H100 GPUs (80GB)

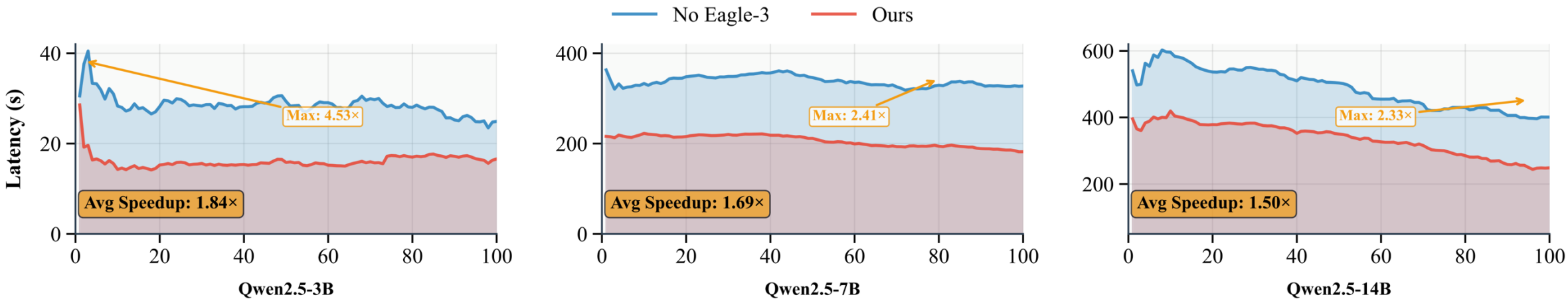
Evaluation

- Stability
  - Measured by validation score during the RL updates
  - Our validation score trajectory remains aligned with the no-acceleration baseline even after 400 steps, while EAGLE-3 often causes degradation, e.g., dropping to 0.15 after step 400



Evaluation

- Speedup
  - Efficiency, measured by wall-clock speedup relative to the baseline training without acceleration.
  - Our approach yields up to 4.53× maximum end-to-end training speedup, with an average improvement of around 1.84×



Conclusion

ReSpec

Rollout ***Adaptive Speculative Decoding***

Dynamically switches between speculative and non-speculative execution modes according to the current system state.

Inference

Training ***Online Learner***

Maintains the draft model through online training by on-policy signals and knowledge distillation.

qiaoling.chen@ntu.edu.sg