FAISys 2025

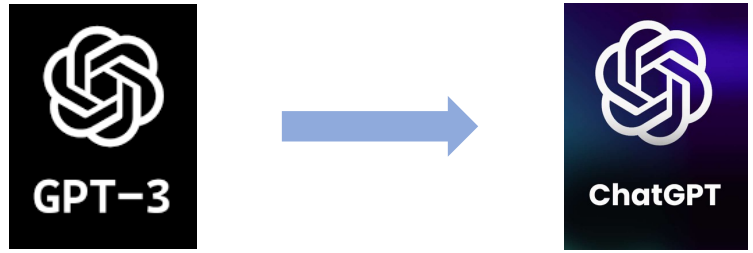
Efficient RL for LLMs with Dynamic and Online Speculative Decoding

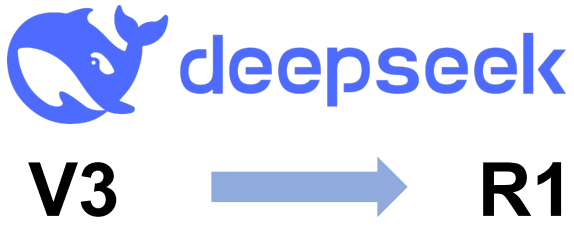
Chao Jin[1], Yinmin Zhong[1], Zili Zhang[1], Yimin Jiang[2], Yibo Zhu[3]

[1]*Peking University* [2]*Anuttacon* [3]*StepFun*

Why use Reinforcement Learning for LLMs?

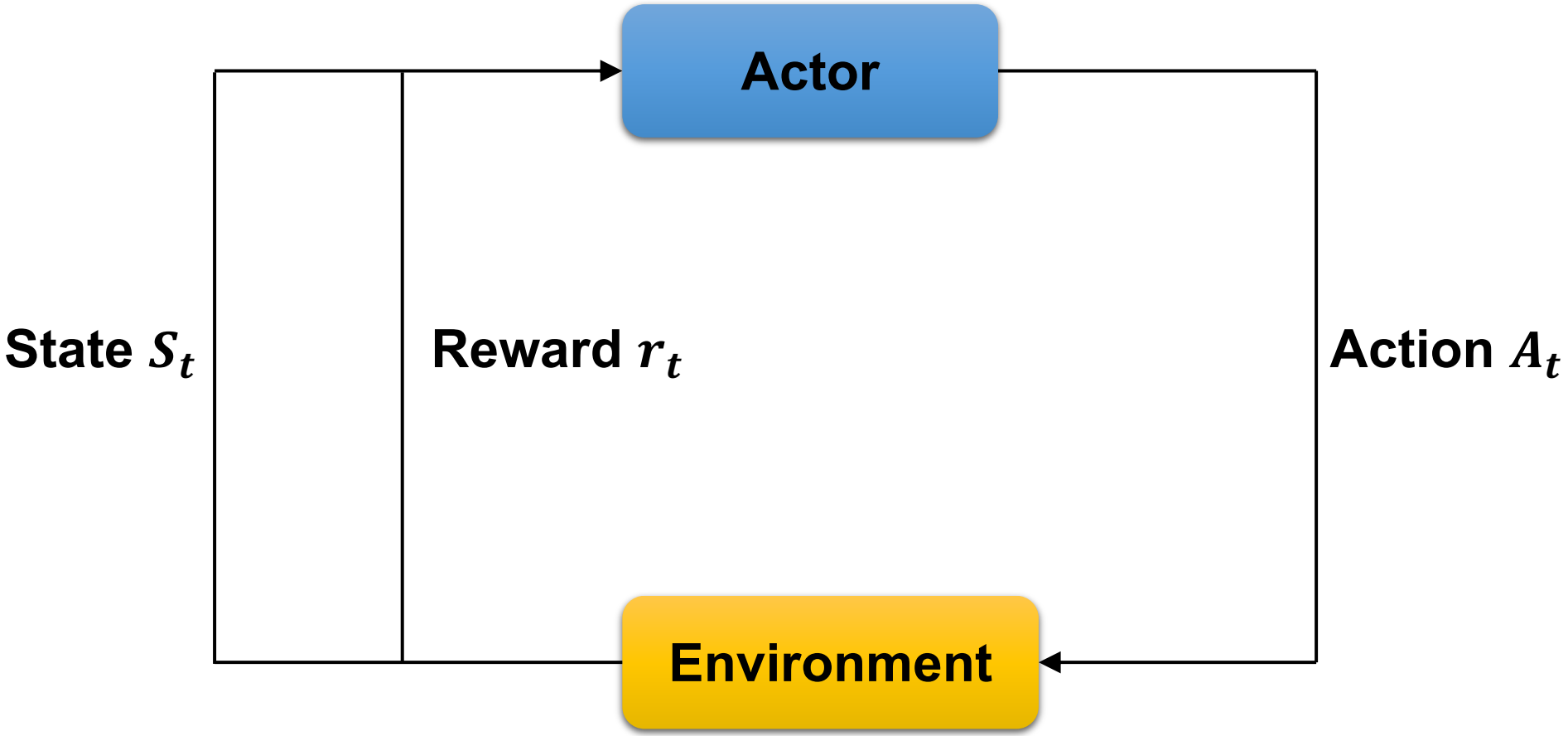
- Human preference alignment
- Safety and robustness
- Hallucination alleviation
- Reasoning (math, STEM)
- Agentic abilities for complex tasks
- …



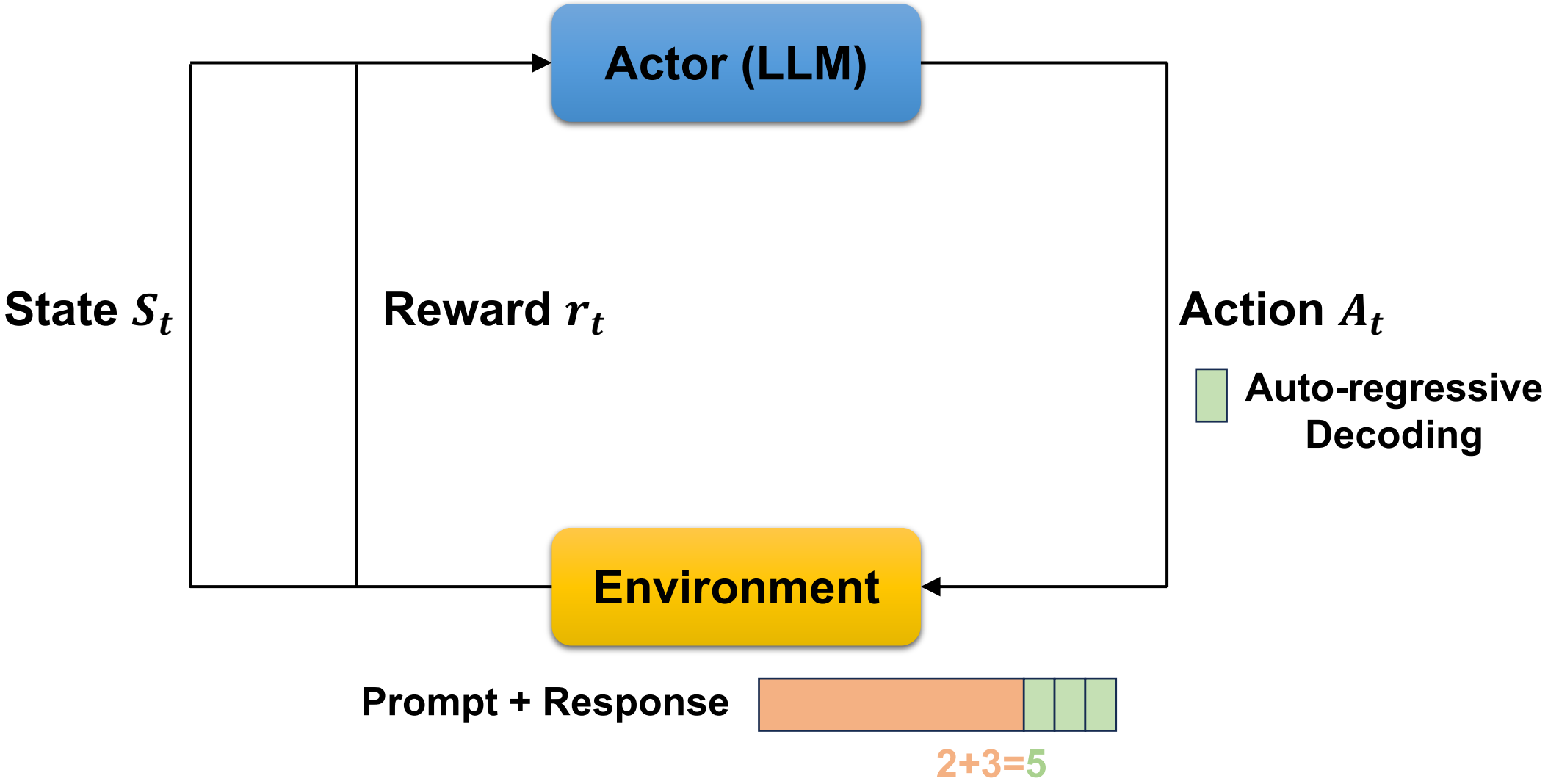


V3 → **R1**

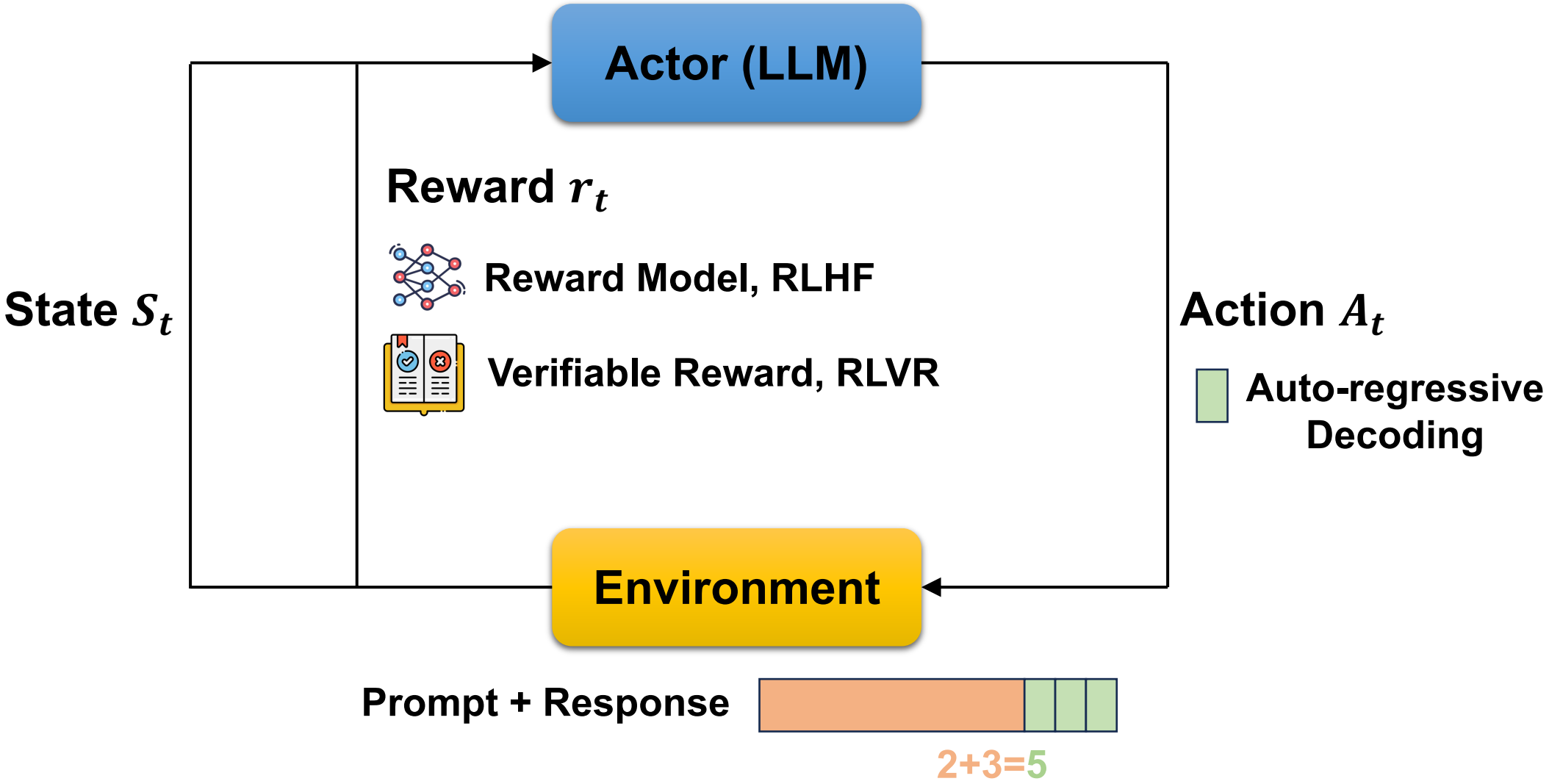
What is RL for LLMs?



What is RL for LLMs?

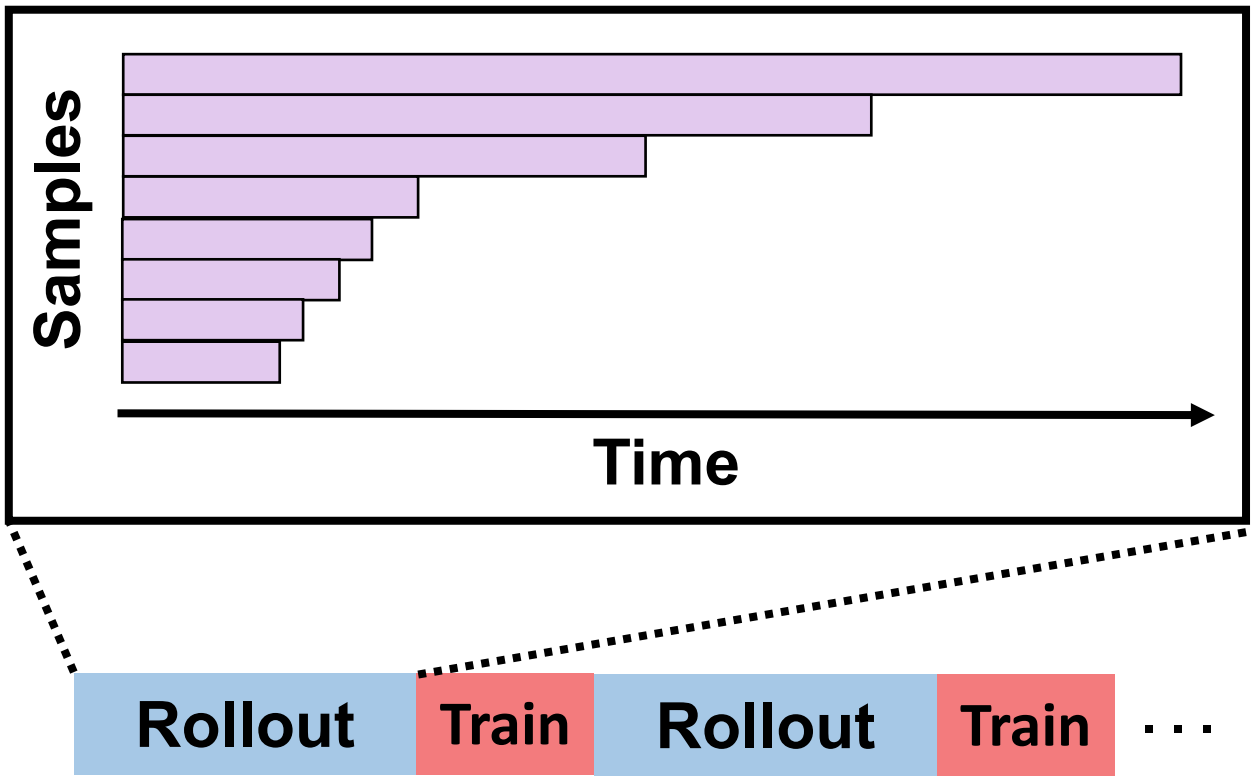


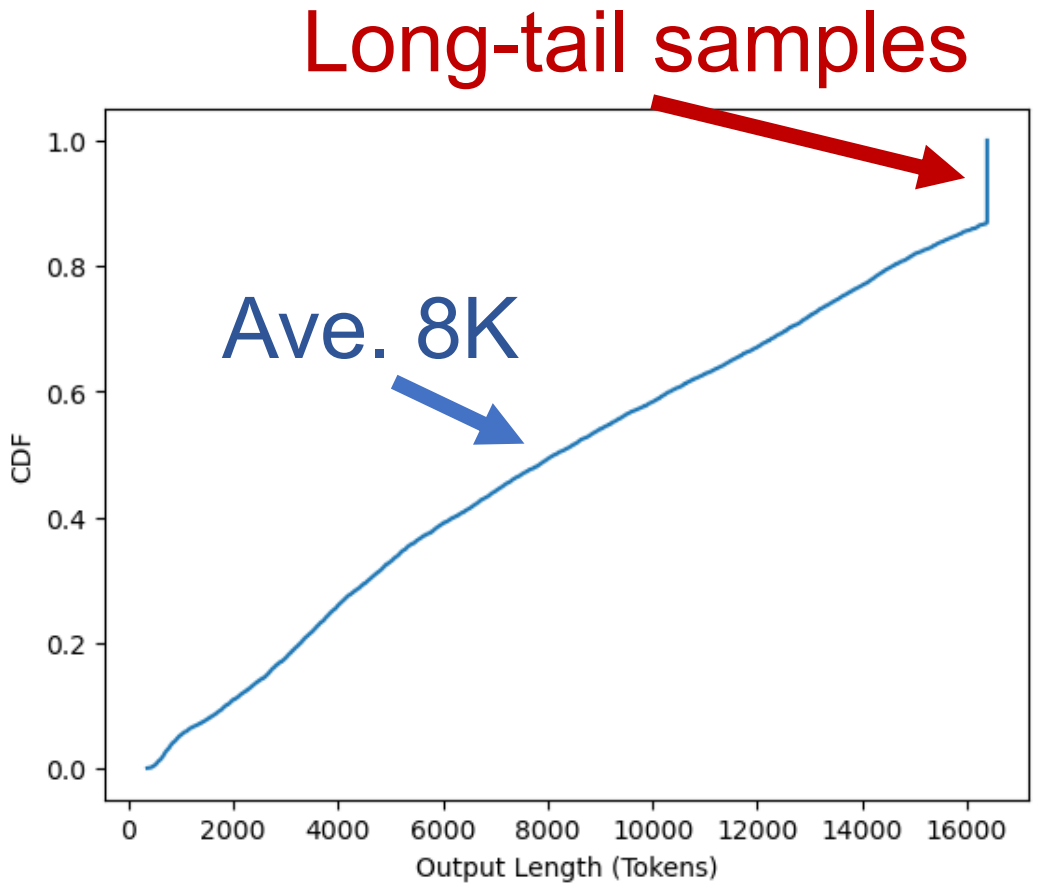
What is RL for LLMs?



Challenge： long-tail phenomenon

Long-tail distribution of output lengths (i.e., # decoding iterations)





Possible solutions for long-tail rollout acceleration

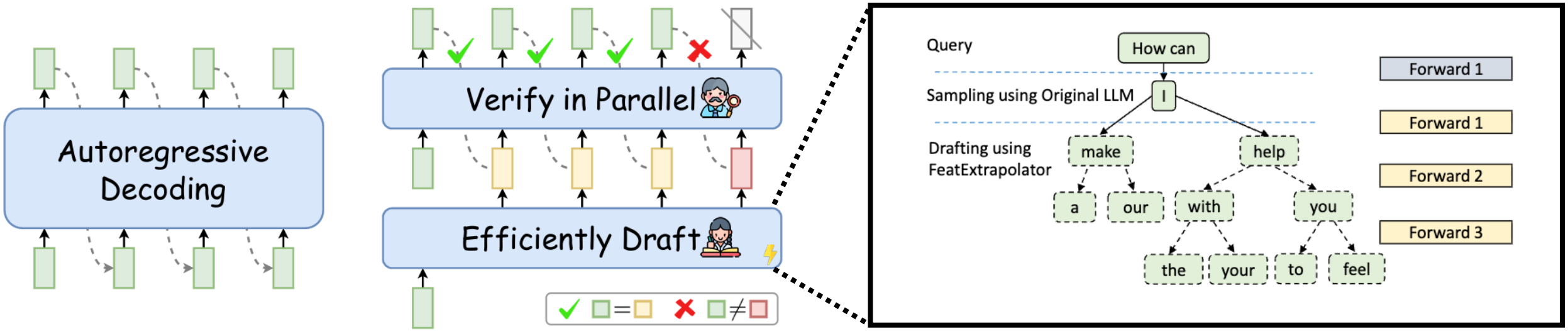
Core problem: How to accelerate a small BS decoding with over-provided resources?

- Larger Tensor Parallelism
  - Improve aggregated memory bandwidth 😀
  - High communication overhead ☹️

- Isolated long-tail generation
  - Isolate long-tail samples to dedicated resources 😀
  - Can long-tail samples be identified? 🧐

Our choice: Speculative Decoding (SD)

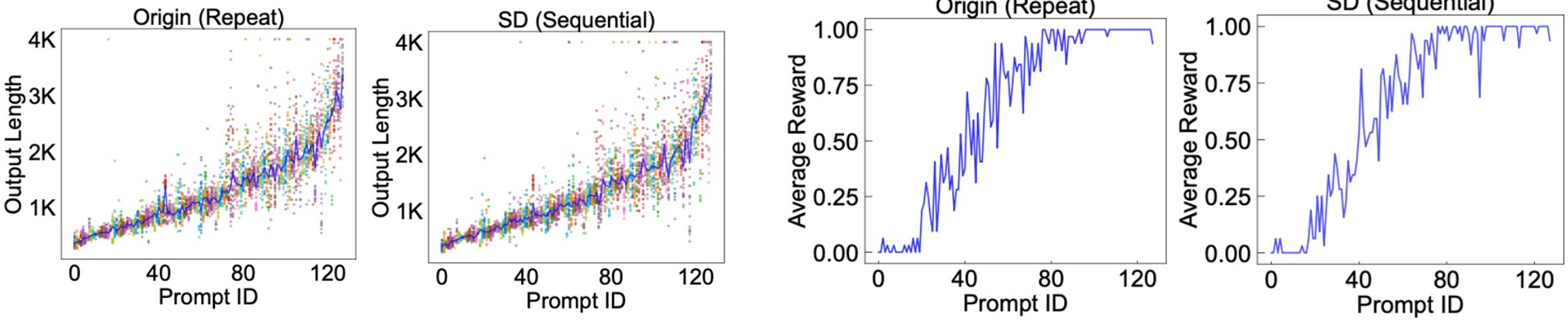
1. Alleviate the memory-bandwidth-bounded problem



[2] EAGLE: Speculative Sampling Requires Rethinking Feature Uncertainty. In ICML'24.

Our choice: Speculative Decoding (SD)

2. Algorithmically equivalent to auto-regressive decoding



When apply Speculative Decoding to RL Rollout …

	Inference Serving	RL Rollout
Data	Generic	Task/Domain-Specific
Model	Fixed Parameters	Continuously Updating Parameters
System	Small batch size (BS)	From large BS to small BS as the rollout stage progresses

SpecRL Overview

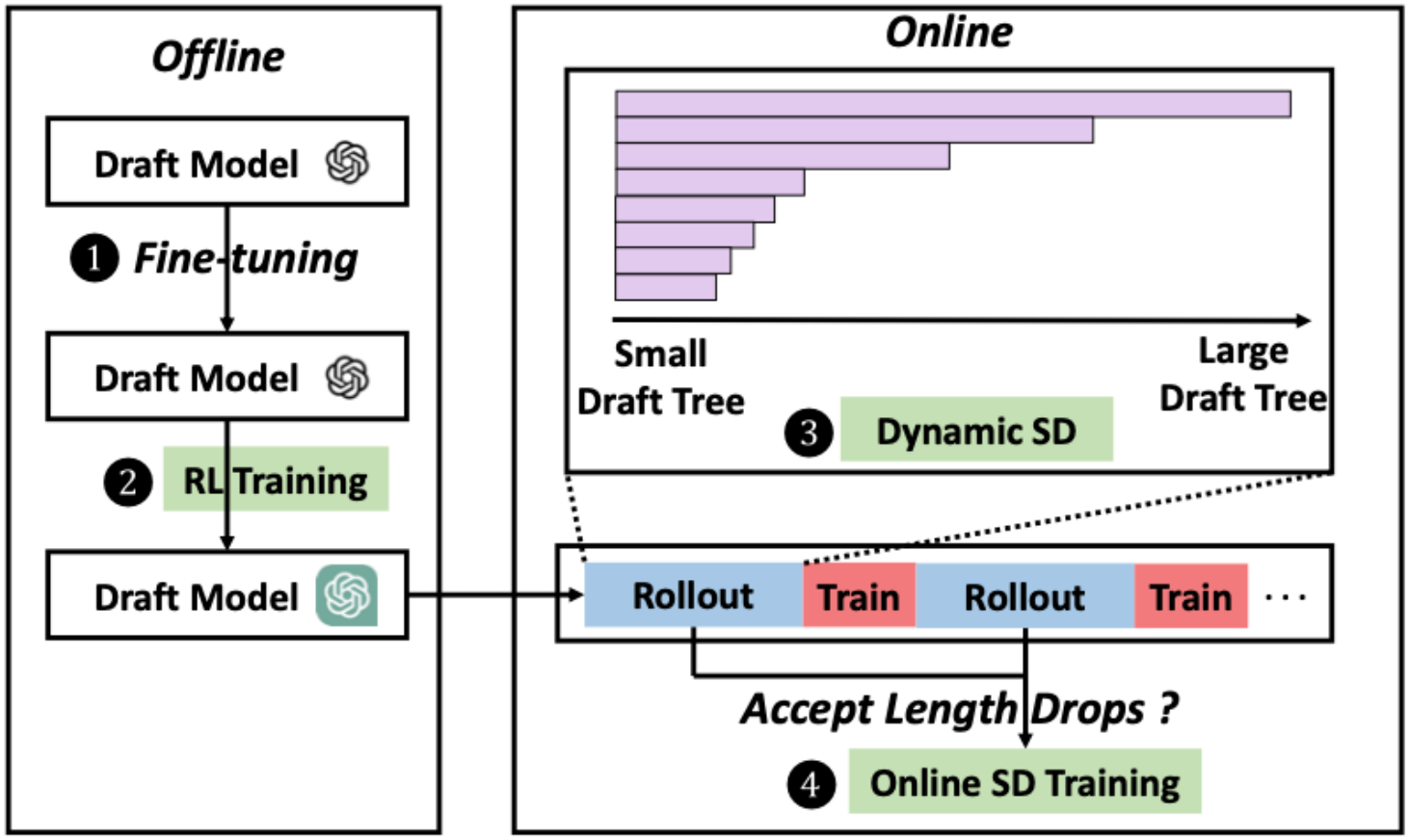


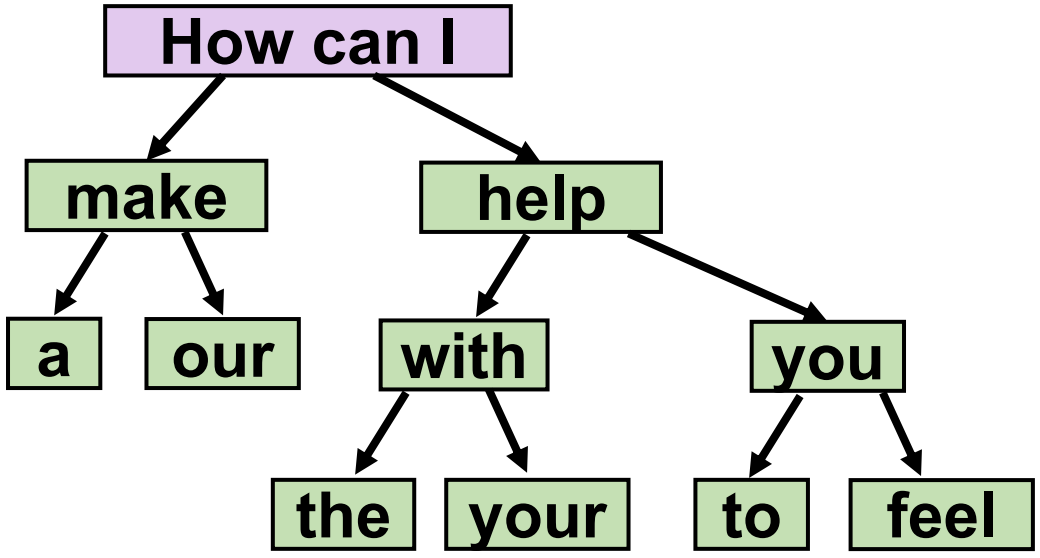
Figure 1: Overview of SpecRL.

Dynamic SD Configurations

Example：

Speculate the next 3 tokens

Select the top-2 most likely tokens to expand



Dynamic SD Configurations

Speculate the next 3 tokens

Select the top-2 most likely tokens to expand

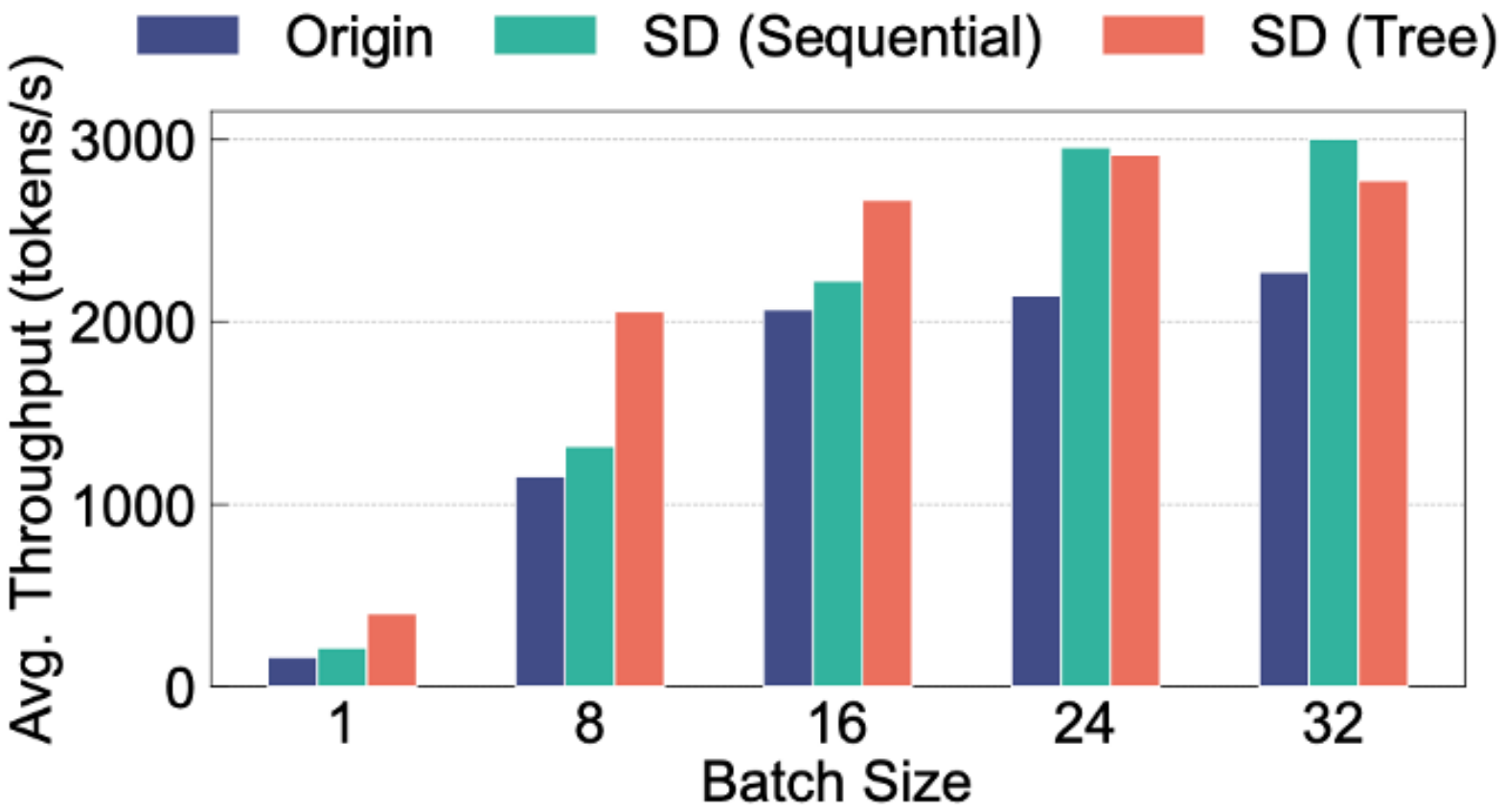
3 **draft iterations** with a **batch size** of 2

- Larger means more speculation
  - Larger accept length, more tokens likely to be accepted 😀
  - More wasted computation ☹️

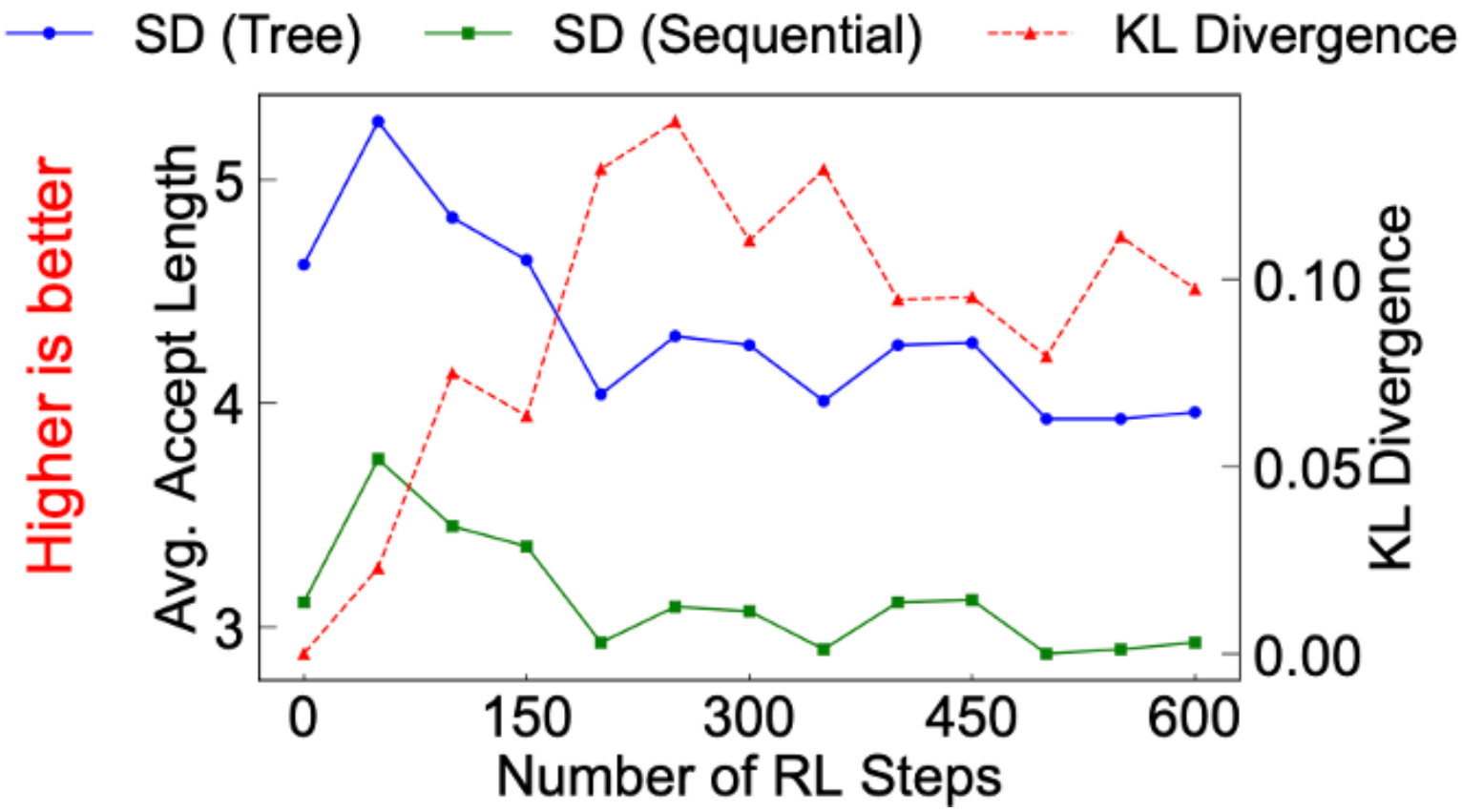
Sequential for large BS: 1 draft iteration with top-1 selection

Tree for small BS: 5 draft iterations with top-8 selection

Preliminary Results

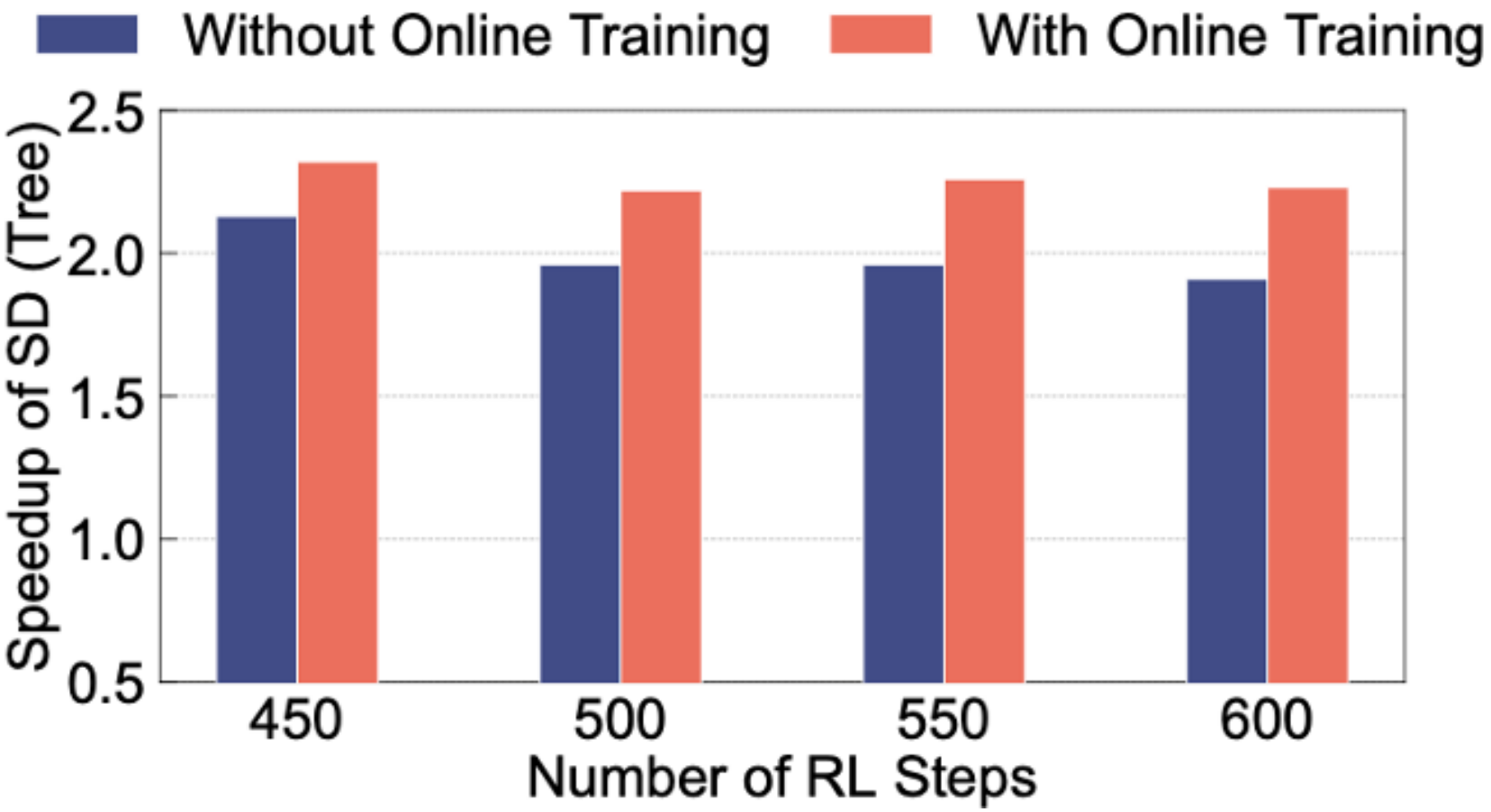


Online Draft Model Training



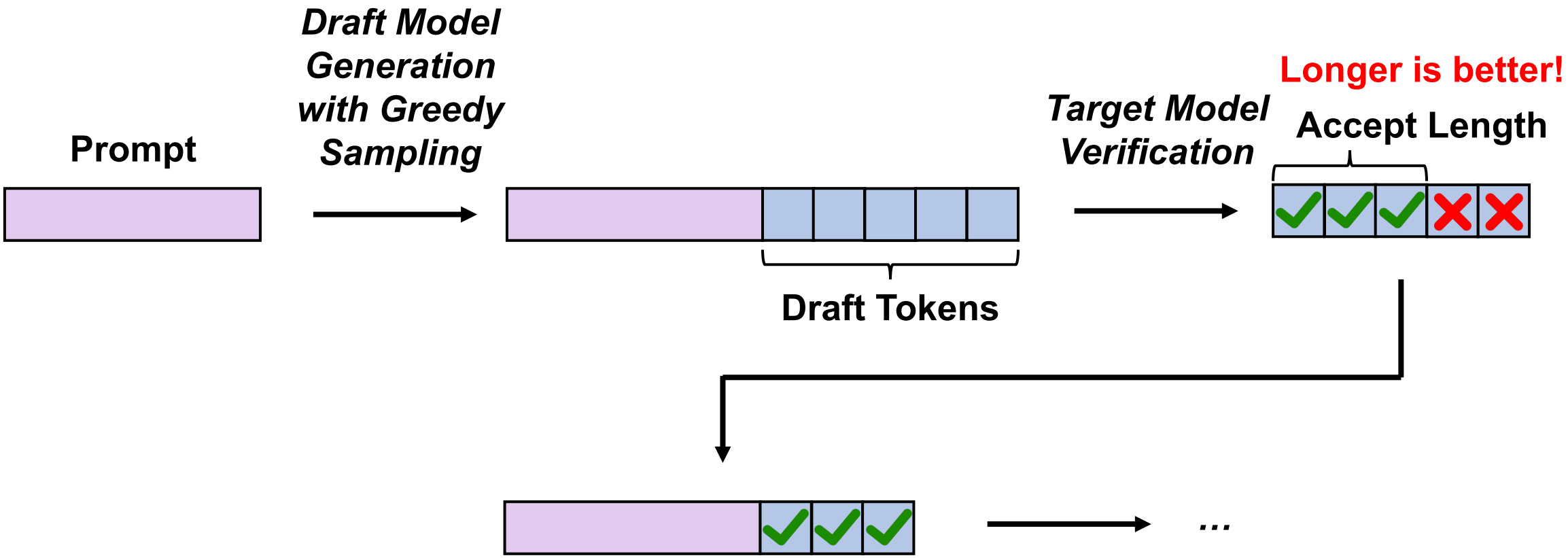
RL steps “slowly”: periodically tuning the draft model ~ 100 RL steps.

Preliminary Results



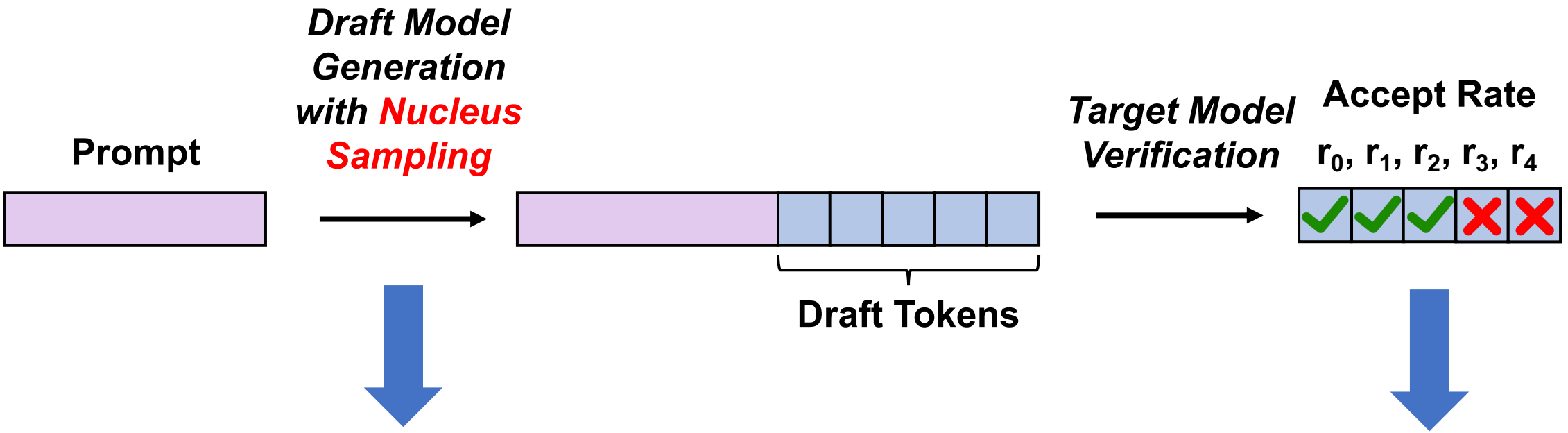
RL training for draft model

How to improve the SD speedup further?



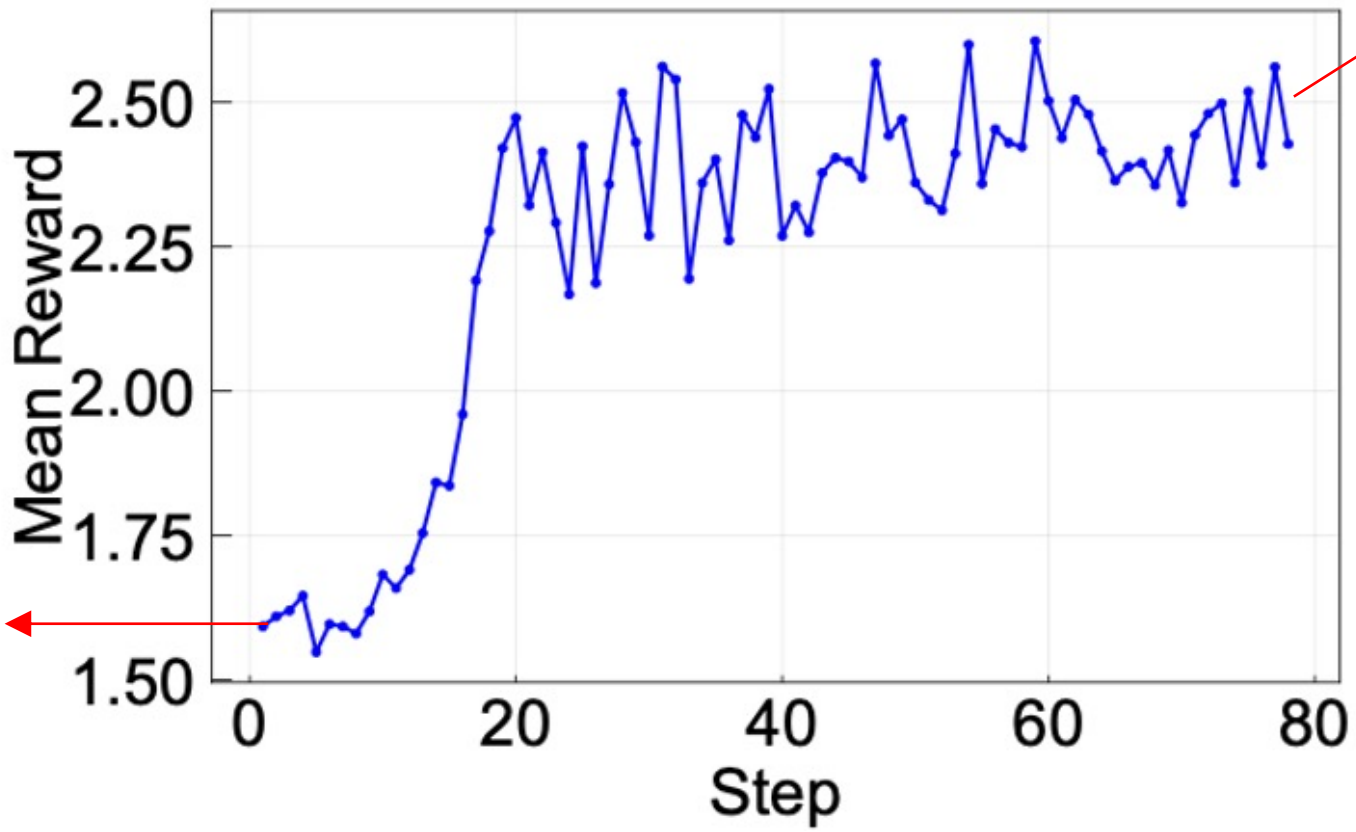
RL training for draft model

Apply RL to SD training to accelerate RL!



Preliminary Results

Accept Length = 6.4



Accept Length = 4.27, which is saturated after typical SD training

Conclusion

- SpecRL: Accelerating the long-tail rollout stage in RL for LLMs
  - **Dynamic SD configuration** that adapts to dynamic batch sizes
  - **Online training** to align the draft model with the evolving target model
  - Domain-specific RL algorithm to improve the draft model

Thanks!


chaojin@pku.edu.cn