

SPRINT: Enabling Interleaved Planning and Parallelized Execution in Reasoning Models

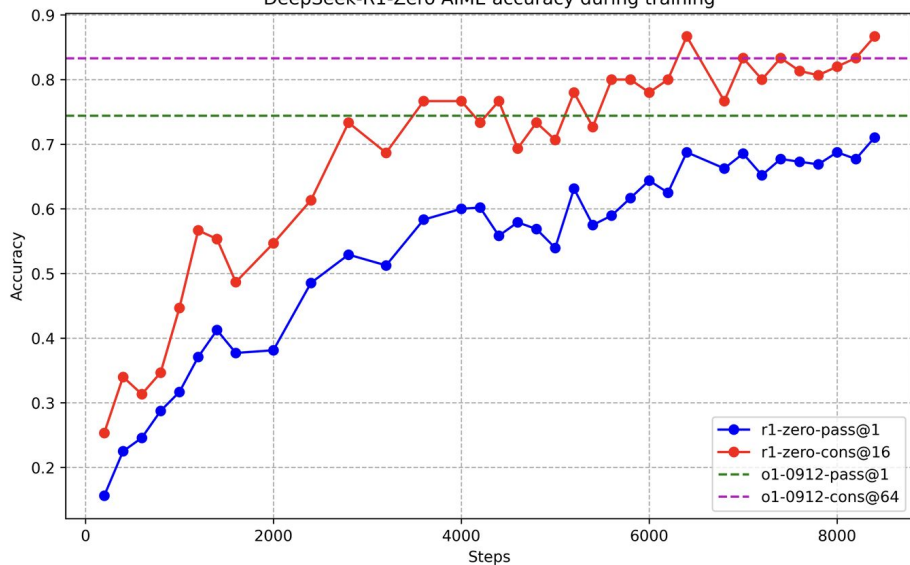
Emil Biju*, Shayan Talaei*, Zhemin Huang*, Mohammadreza Pourreza
Azalia Mirhoseini, Amin Saberi

Stanford University, Google

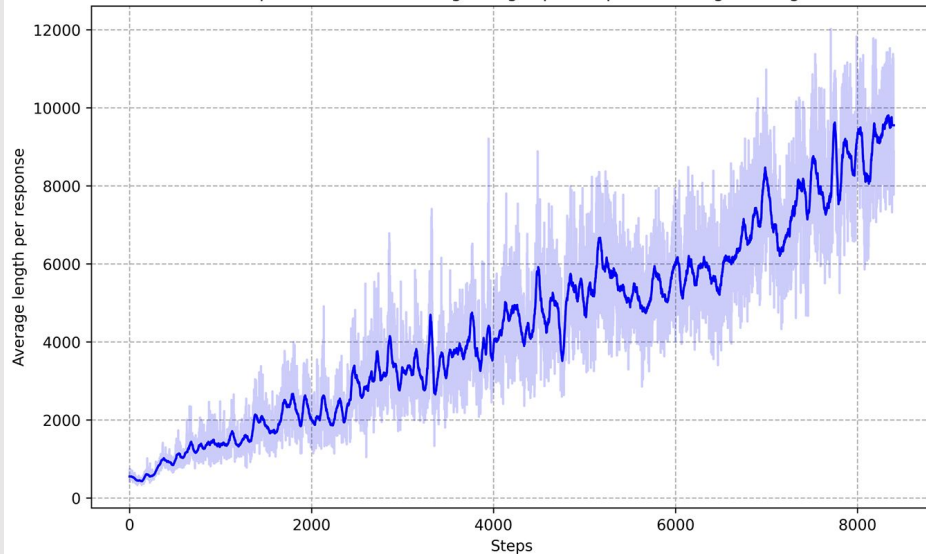
Large Reasoning Models

Long sequential Chains-of-Thoughts, Higher accuracy in the reasoning tasks

DeepSeek-R1-Zero AIME accuracy during training



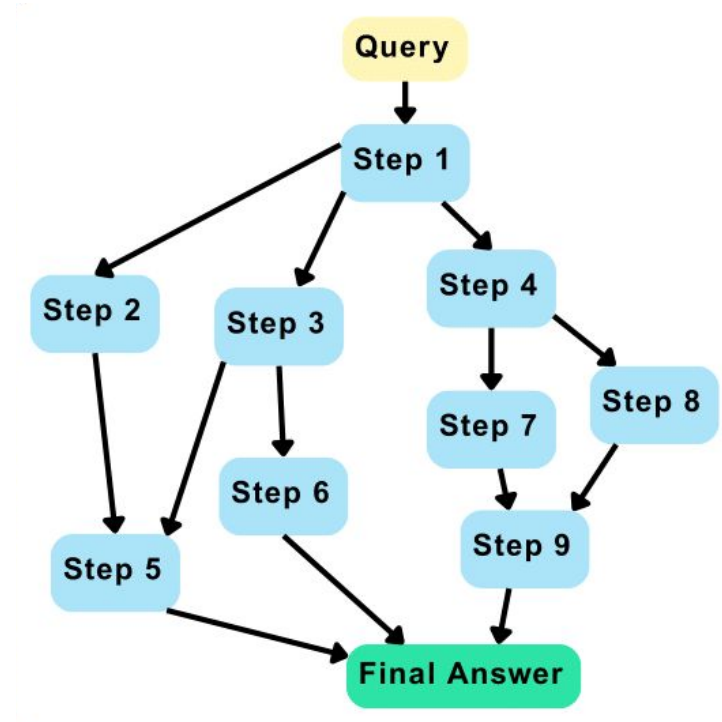
DeepSeek-R1-Zero average length per response during training



DeepSeek-R1, DeepSeek-AI (2025)

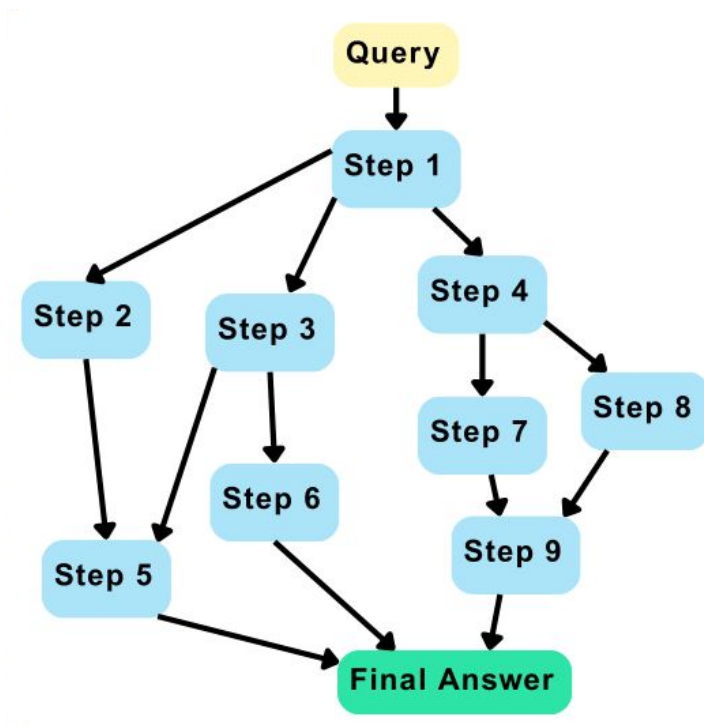
Does Reasoning Have to Be Sequential?

- 🔍 **Observation:** Many reasoning steps appear to be independent of one another and could therefore be parallelized, e.g.
 - Trying alternative approaches
 - Decomposing a task into subtasks
 - Verifying previous reasoning steps



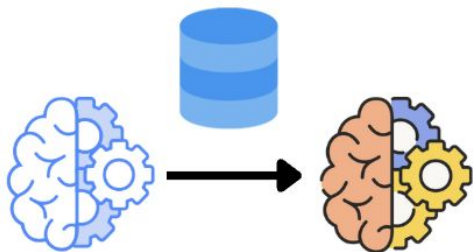
Does Reasoning Have to Be Sequential?

- 🔍 **Observation:** Many reasoning steps appear to be independent of one another and could therefore be parallelized, e.g.
 - Trying alternative approaches
 - Decomposing a task into subtasks
 - Verifying previous reasoning steps
- 💡 **Idea:** Get the model to identify the parallelization opportunities during the reasoning process



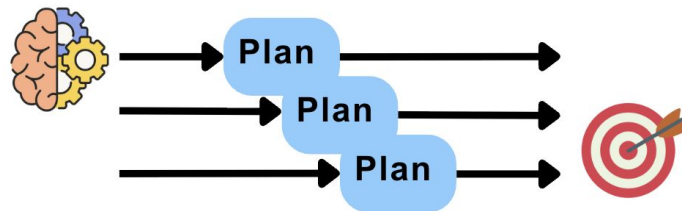
SPRINT Framework

Post-Training Recipe



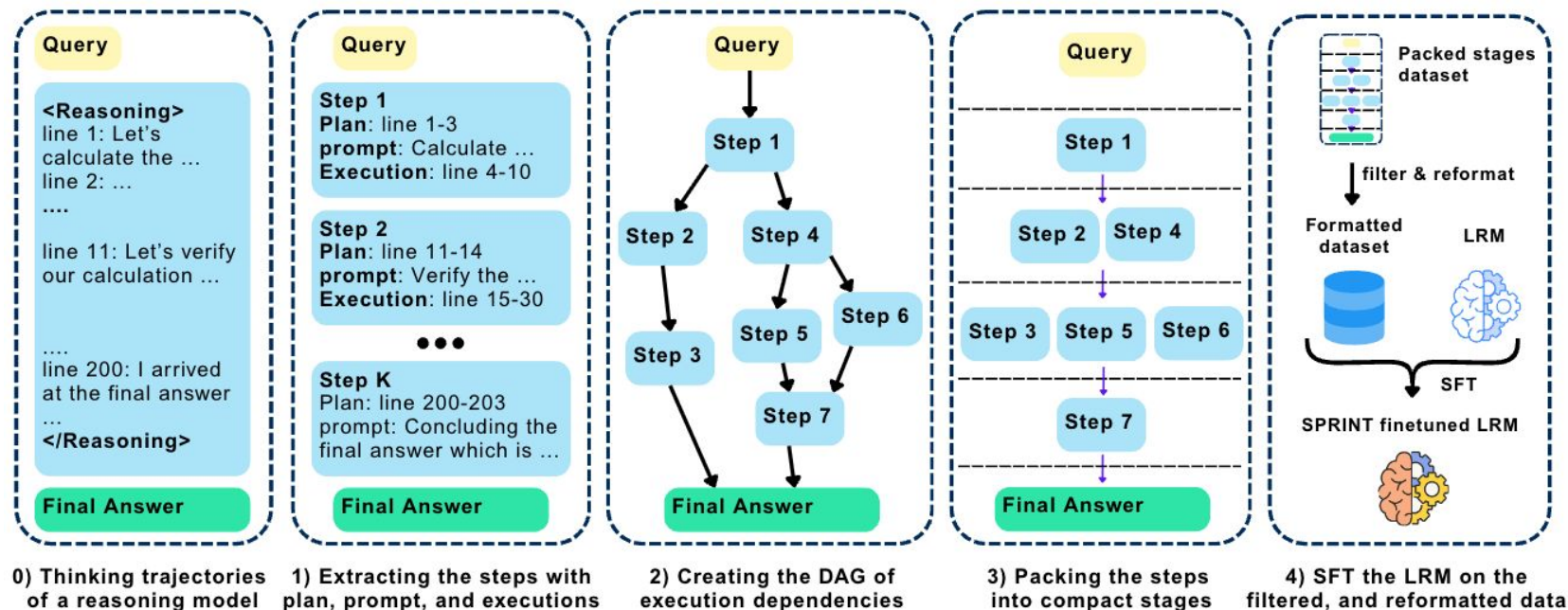
Teach the model how to identify the parallelization opportunities

Inference

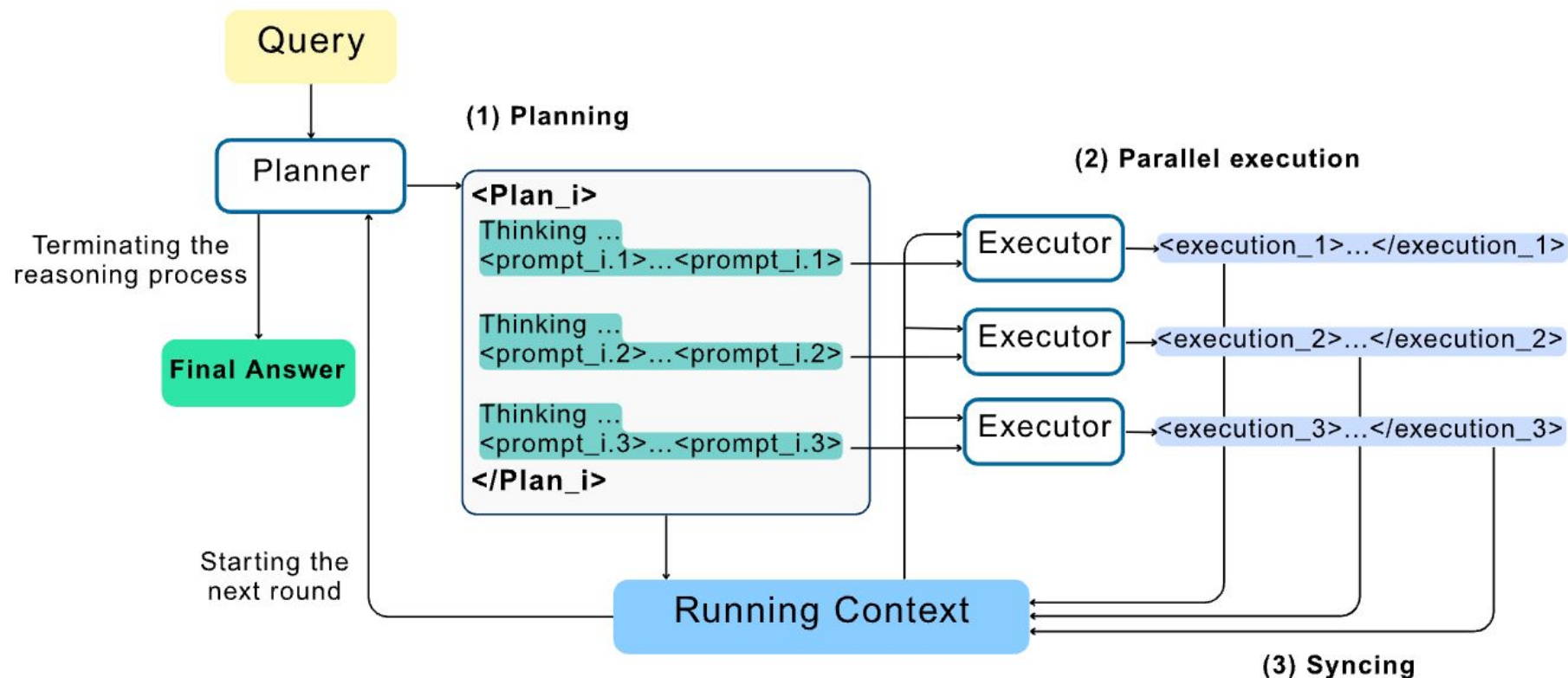


Propose independent plans and execute them in parallel

SPRINT Post-Training Process

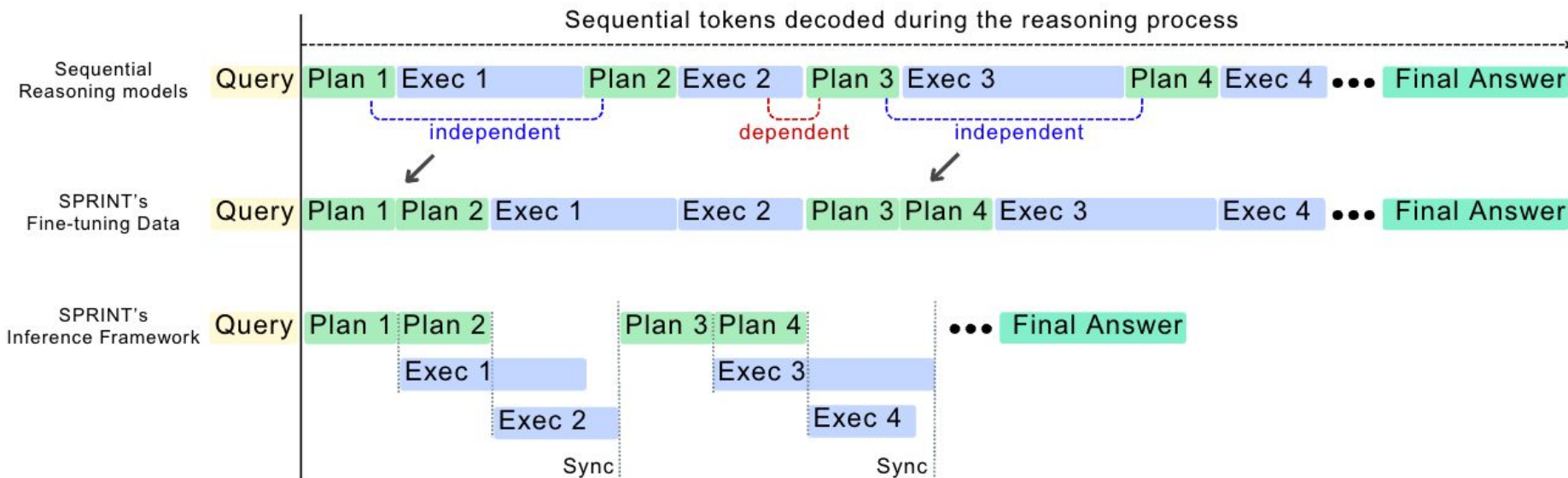


SPRINT Inference Process





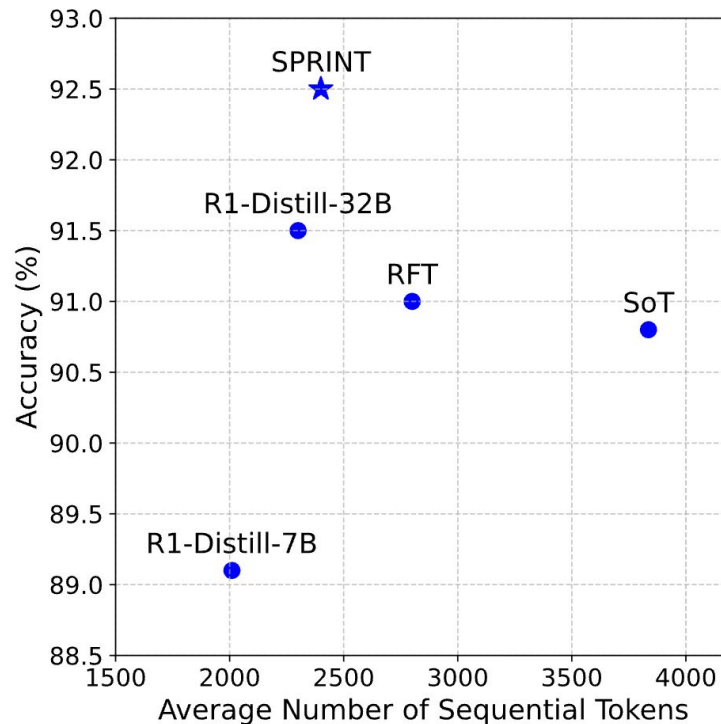
Overlapping Execution Steps

SPRINT would save on the **sequential decoding steps** by overlapping **concurrent execution** of **independent plans**.



Higher Accuracy with Fewer Decoding Steps

-  **Training recipe**
 - Starting from 6k Deepseek-R1 thinking trajectories on MATH train set
 - Filtering the samples with low parallelization opportunities
 - SFT DeepSeek-R1-Distill-Qwen-7B on the reformatted reasoning trajectories
-  **Results**
 - On MATH test set SPRINT improves the pareto frontier of accuracy vs seq. tokens
 - Beating x4 larger model

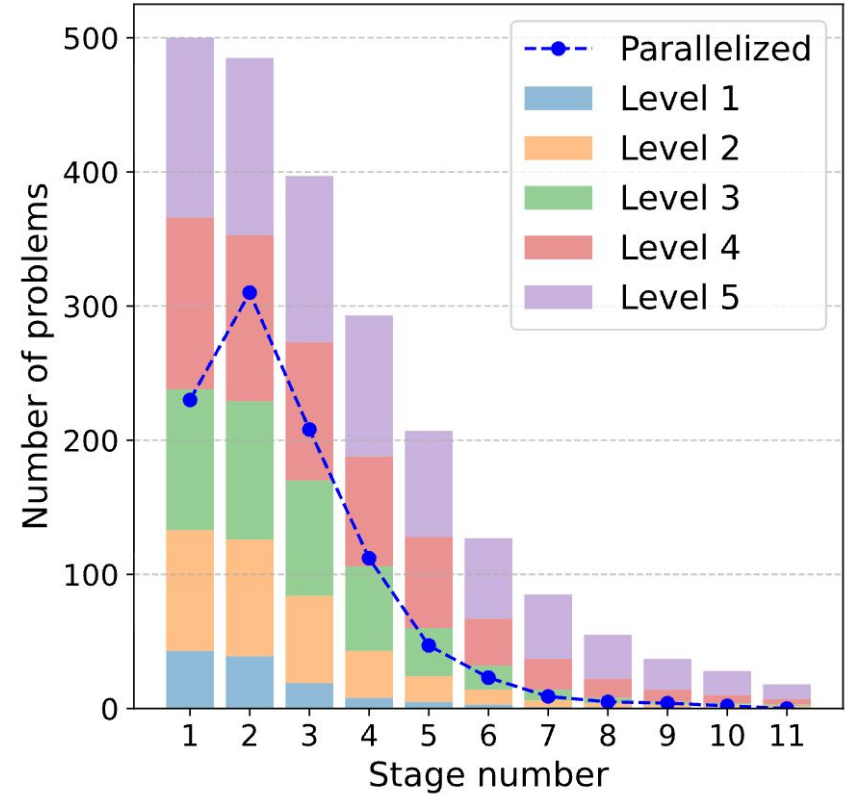


Generalization Beyond the Training Domain

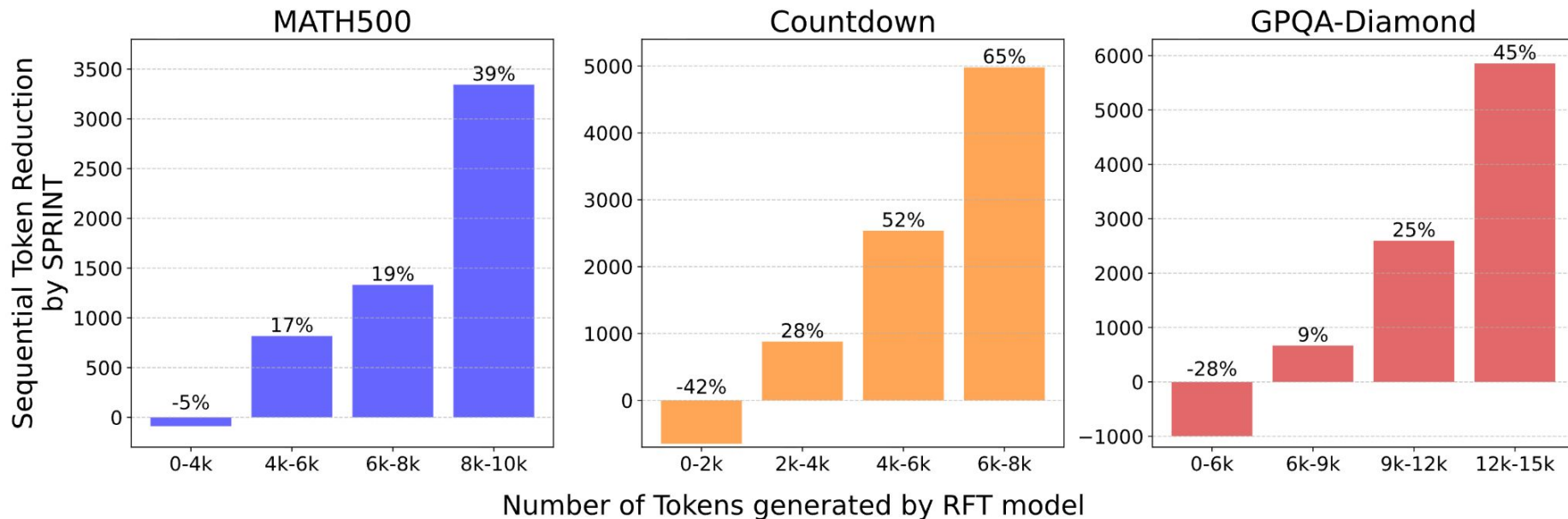
Method	In-domain		Out-of-domain			
	MATH500		Countdown		GPQA-Diamond	
	Acc↑	Seq. Tokens↓	Acc↑	Seq. Tokens↓	Acc↑	Seq. Tokens↓
Self-consistency	80.5	590	78.5	2845	45.4	4735
SoT-chat	47.3	256	80.0	2367	49.4	3526
SoT-reasoning	90.8	3836	82.4	5823	48.0	7560
RFT	91.0	2880	84.9	4917	50.5	7103
SPRINT	92.5	2440	85.9	2284	51.0	6336

Parallelization Patterns by SPRINT

- Harder problems require more iterative planning and execution.
- SPRINT introduces more parallelism in the early stages.
- Early parallelism explores multiple approaches before converging to a reliable solution.



Even Larger Relative Saving for Longer Sequences



Next Steps

- **Improving beyond data quality** → Use GRPO to let the model find parallelization strategies beyond supervised samples.
- **Parallelizing tool-use in reasoning models** → Overlap independent, time-consuming tool calls through parallelized planning.
- **Realizing the wall-clock speedups** → Employ hardware-optimized implementations that convert sequential token reduction into lower latency.



Questions?