PRINCETON UNIVERSITY

Carnegie Mellon University

# SpecReason: Fast and Accurate Inference-Time Compute via Speculative Reasoning

Rui Pan, Yinwei Dai, Zhihao Zhang, Gabriele Oliaro, Zhihao Jia, Ravi Netravali

NEURAL INFORMATION PROCESSING SYSTEMS

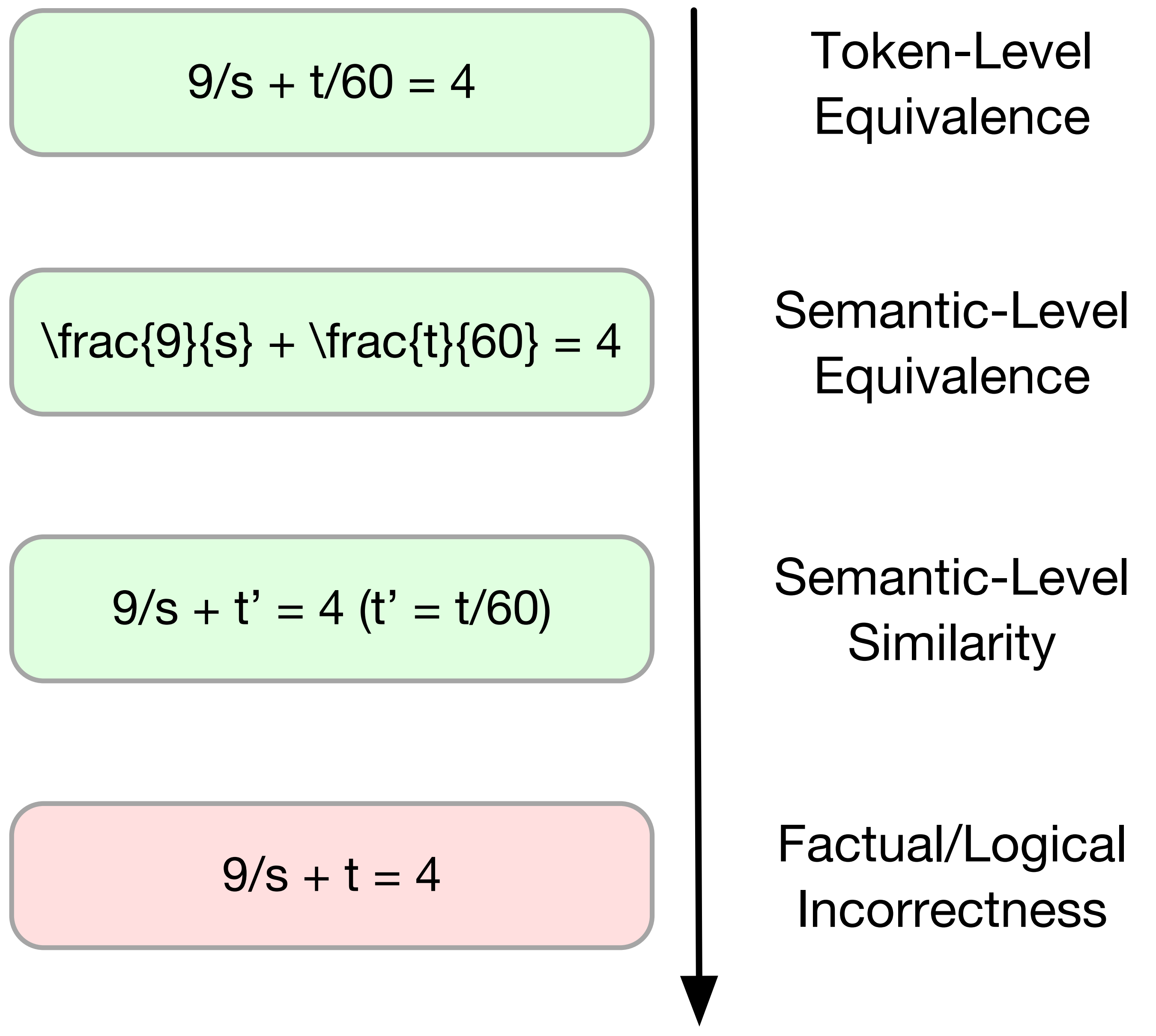
**TL;DR: Speculating reasoning steps for semantic, not token, equivalence speeds up LRM inference by up to 3x with no accuracy loss!**

## Problem

Large Reasoning Models generate thinking tokens in long chains of thought, leading to high inference latency that scales linearly with CoT length (often thousands of tokens).

## Insights

1. Many intermediate steps are easier than end-to-end reasoning (arithmetic, case checks, routine logic)
2. Reasoning progress depend on insights, not exact tokens
3. Occasional mistakes can be corrected via self-reflection



9/s + t/60 = 4

Token-Level Equivalence

\frac{9}{s} + \frac{t}{60} = 4

Semantic-Level Equivalence

9/s + t' = 4 (t' = t/60)

Semantic-Level Similarity

9/s + t = 4

Factual/Logical Incorrectness

The spectrum of approximations of an example reasoning step

Question: Every morning Aya goes for a \$9\$-kilometer-long walk and stops at a coffee shop afterwards. When she walks at a constant speed of \$s\$ kilometers per hour, the walk takes her 4 hours, including \$t\$ minutes spent in the coffee shop. When she walks \$s+2\$ kilometers per hour, the walk takes her 2 hours and 24 minutes, including \$t\$ minutes spent in the coffee shop. Suppose Aya walks at \$s+ \frac{1}{2}\$ kilometers per hour. Find the number of minutes the walk takes her, including the \$t\$ minutes spent in the coffee shop.

Small Model🐇❌

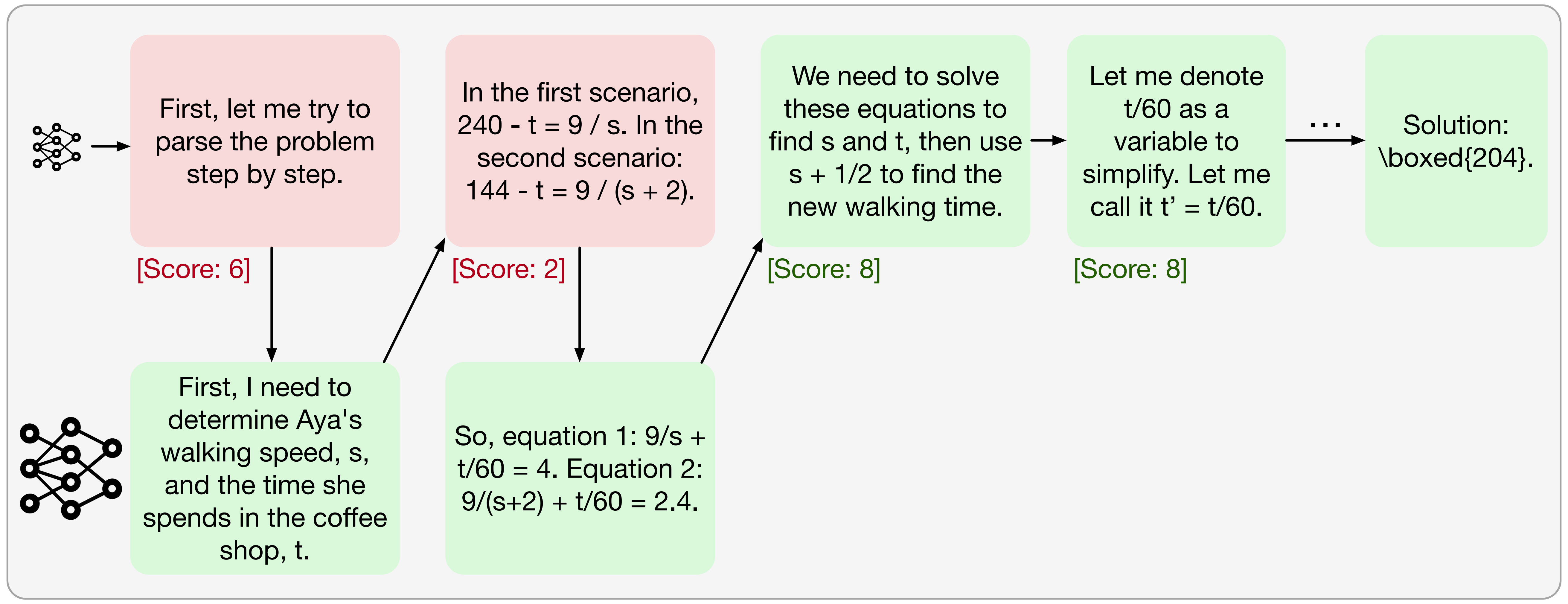
First, let me try to parse the problem step by step. ... In the first scenario, 240 - t = 9 / s. ... Taking the positive root: s ≈ 0.0897 km/h. ... Thus, t = 240 - (48 - 12*sqrt(19)) = 192 + 12*sqrt(19) ≈ 139.6932. ... The result is \boxed{1055} minutes.

Base Model🐢✅

First, I need to determine Aya's walking speed, s, and the time she spends in the coffee shop, t. ... So, equation 1: 9/s + t/60 = 4. Equation 2: 9/(s+2) + t/60 = 2.4. ... Thus, s = 2.5, and t = 24. ... The result is \boxed{204} minutes.

SpecReason🐇✅

Soln: s = 2.5 km/hr, t = 24 mins. The result is 9 / (2.5 + 0.5) * 60 + 24 = 204 minutes.



First, let me try to parse the problem step by step.

[Score: 6]

First, I need to determine Aya's walking speed, s, and the time she spends in the coffee shop, t.

In the first scenario, 240 - t = 9 / s. In the second scenario: 144 - t = 9 / (s + 2).

[Score: 2]

So, equation 1: 9/s + t/60 = 4. Equation 2: 9/(s+2) + t/60 = 2.4.

We need to solve these equations to find s and t, then use s + 1/2 to find the new walking time.

[Score: 8]

Let me denote t/60 as a variable to simplify. Let me call it t' = t/60.

[Score: 8]

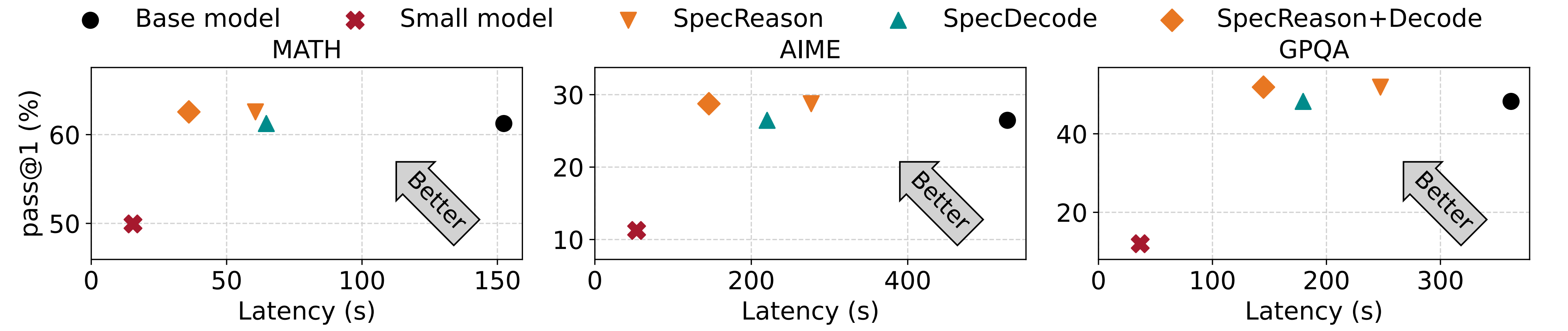
...

Solution: \boxed{204}.

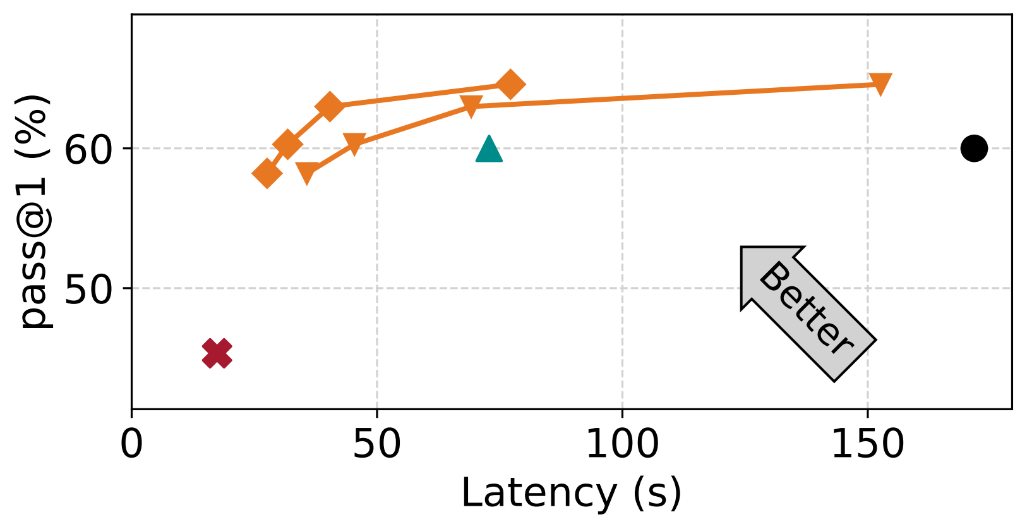
## Method

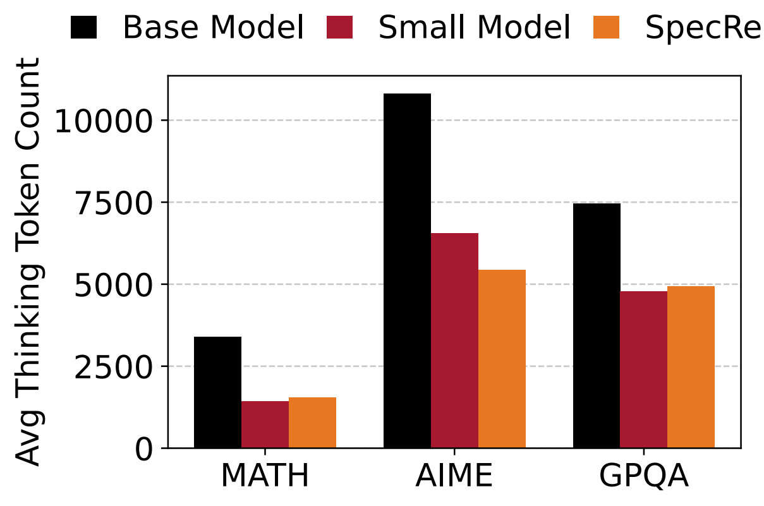
- First generate reasoning steps with a small model; base model "verifies" each step and generates a utility score
- Accept semantically useful steps (utility score > acceptance threshold); otherwise, reject and regenerate with base model
- Verification uses a single prefill pass to generate a single token→extremely low overhead
- Complementary with speculative decoding during regeneration

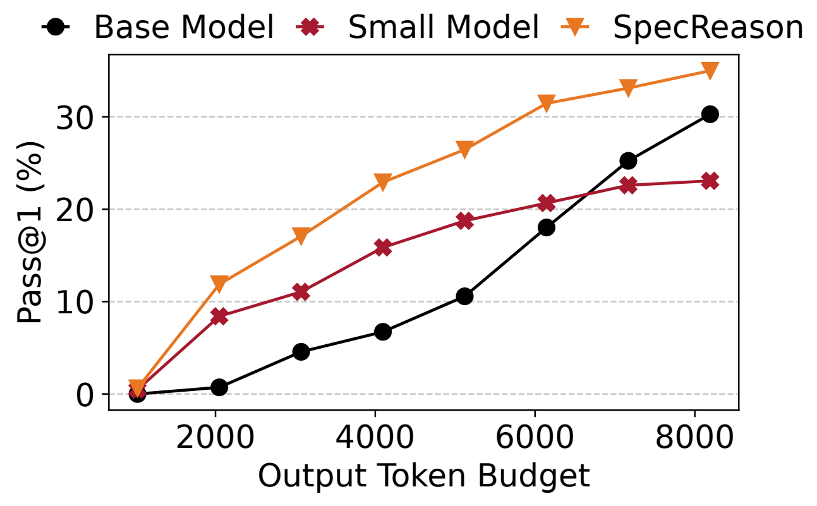
## Results



- Fast (1.4-3.0× speedup) and accurate (+0.4-9.0%) LRM inference!
- Trades off accuracy for latency via changing the acceptance threshold
- Reduces #output tokens, improving accuracy under same token budget


pass@1 (%)
Latency (s)
Better


■ Base Model
■ Small Model
■ SpecReason
Avg Thinking Token Count
MATH
AIME
GPQA


● Base Model
✖ Small Model
▼ SpecReason
Pass@1 (%)
Output Token Budget