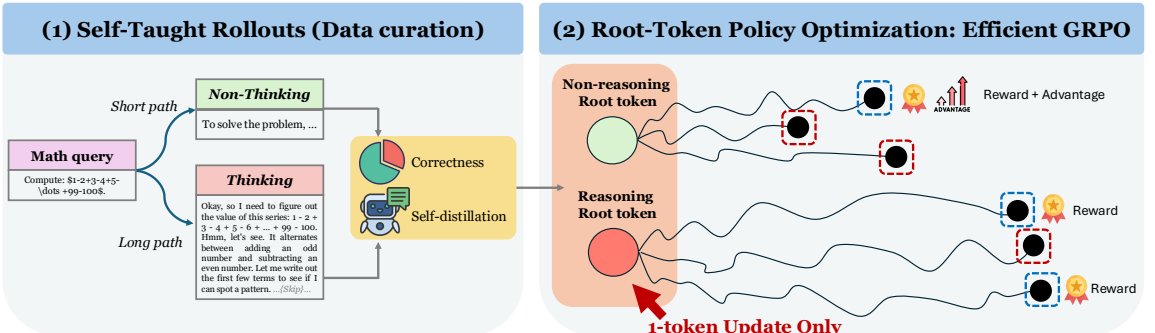

(1) Self-Taught Rollouts (Data curation)
Short path
Non-Thinking
To solve the problem, ...
Math query
Compute: $1-2+3-4+5-\dots +99-100$.
Long path
Thinking
Okay, so I need to figure out the value of this series: 1 - 2 + 3 - 4 + 5 - 6 + ... + 99 - 100. Hmm, let's see. It alternates between adding an odd number and subtracting an even number. Let me write out the first few terms to see if I can spot a pattern. ...(Skip)...
Correctness
Self-distillation
(2) Root-Token Policy Optimization: Efficient GRPO
Non-reasoning Root token
Reasoning Root token
ADVANTAGE
Reward + Advantage
Reward
Reward
1-token Update Only