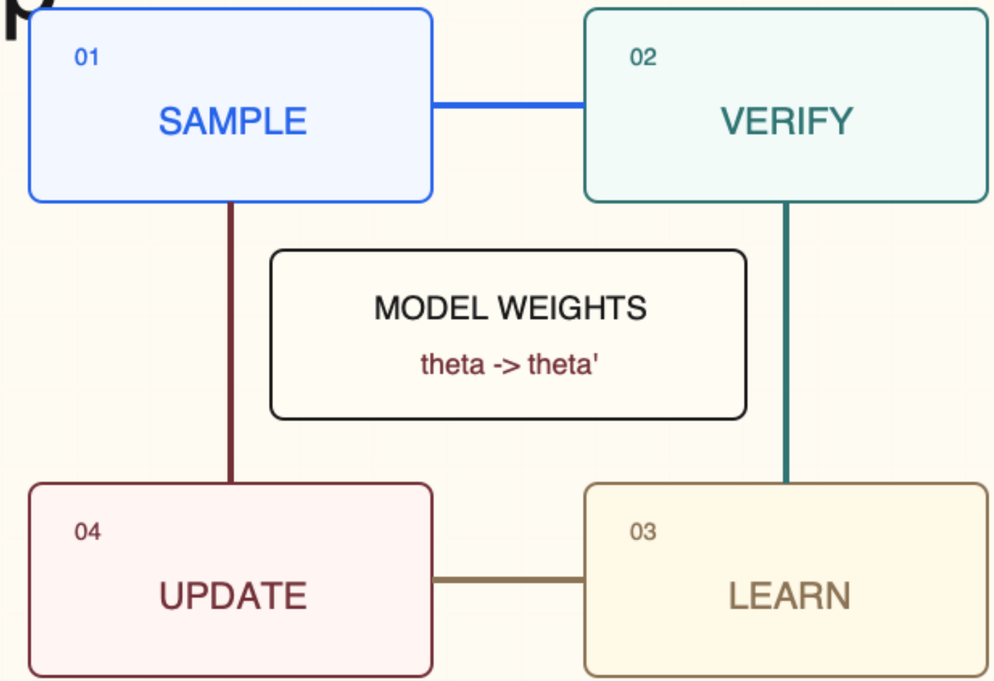
TRAIN-TIME FEEDBACK LOOP

PART 6 00:00-05:43

Train-Time Scaling Closes the Feedback Loop

Stanford CS329A Part 6 - generated data becomes useful only through verification and parameter updates



GENERATE | CHECK | LEARN | CHANGE

COMPUTE REGIMES

PART 6 03:07-07:40 / 11:31-16:08

Keep Four Compute Regimes Separate

They spend compute at different moments, expose different supervision, and do not all change the weights.

01

PRETRAIN

before deployment

WEIGHTS OPEN

changes policy

02

SFT

demonstrations

WEIGHTS OPEN

changes policy

03

TEST-TIME

search / sample

WEIGHTS LOCKED

changes outputs

04

TRAIN-TIME

generated experience

WEIGHTS OPEN

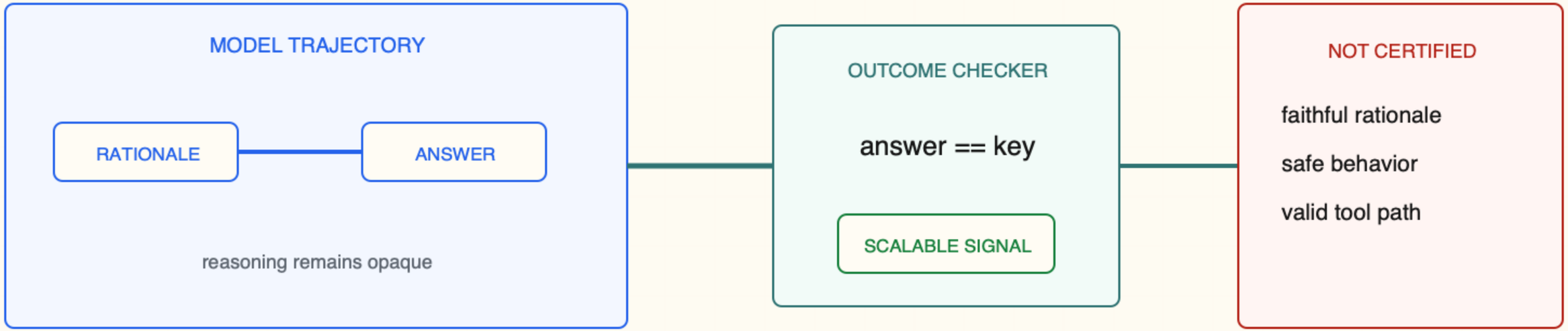
changes policy

COMPUTE LOCATION != LEARNING MECHANISM

FEEDBACK BOTTLENECK

Verifiability Is the Feedback Bottleneck

A correct final answer can create scalable signal without validating the rationale, behavior, or tool path.



CAVEAT

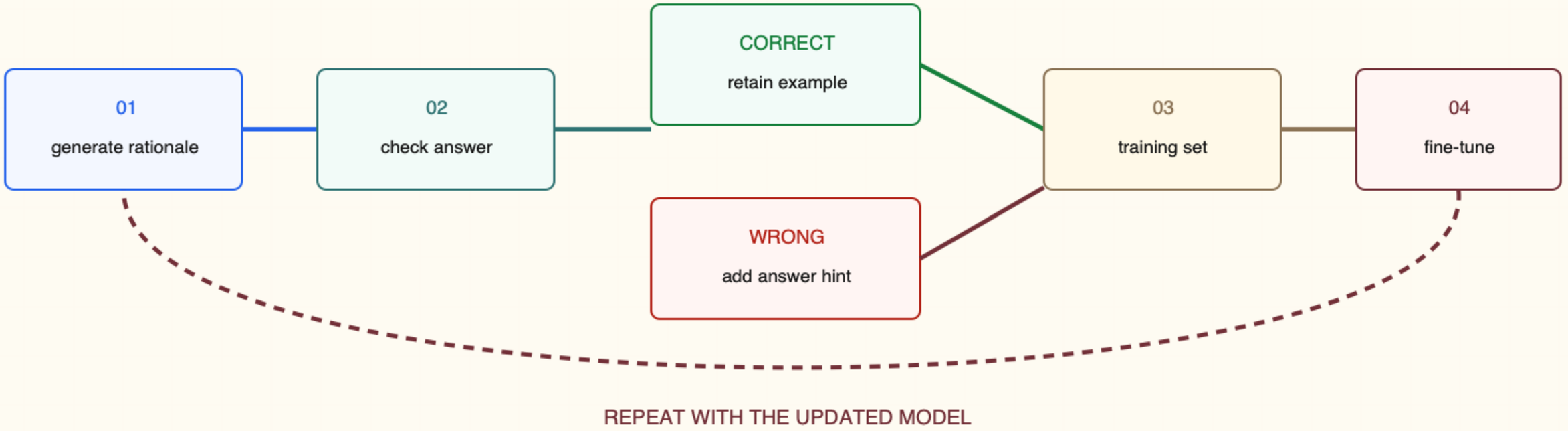
Outcome verification creates useful signal, but it does not certify the process that produced the answer.

STAR BOOTSTRAPPING

PART 6 16:08-27:17

STaR Bootstraps Rationales

Keep correct-answer rationales, rationalize failures with an answer hint, fine-tune, and repeat.

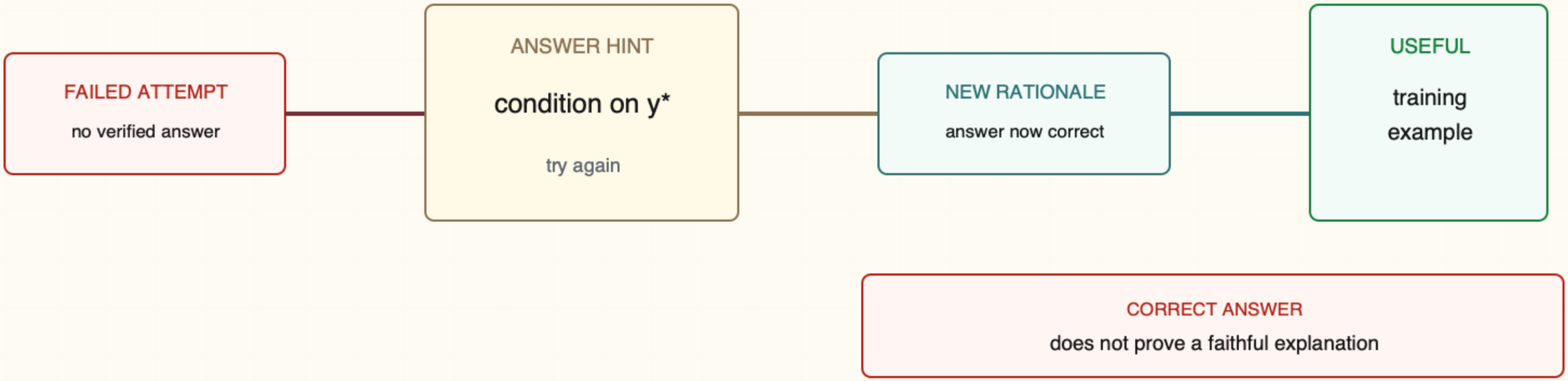


ANSWER-HINT CAVEAT

PART 6 17:53-25:13 / 30:49-33:54

Rationalization Helps and Can Mislead

The answer hint may reveal a usable path or invite a plausible post-hoc story that the verifier never checked.

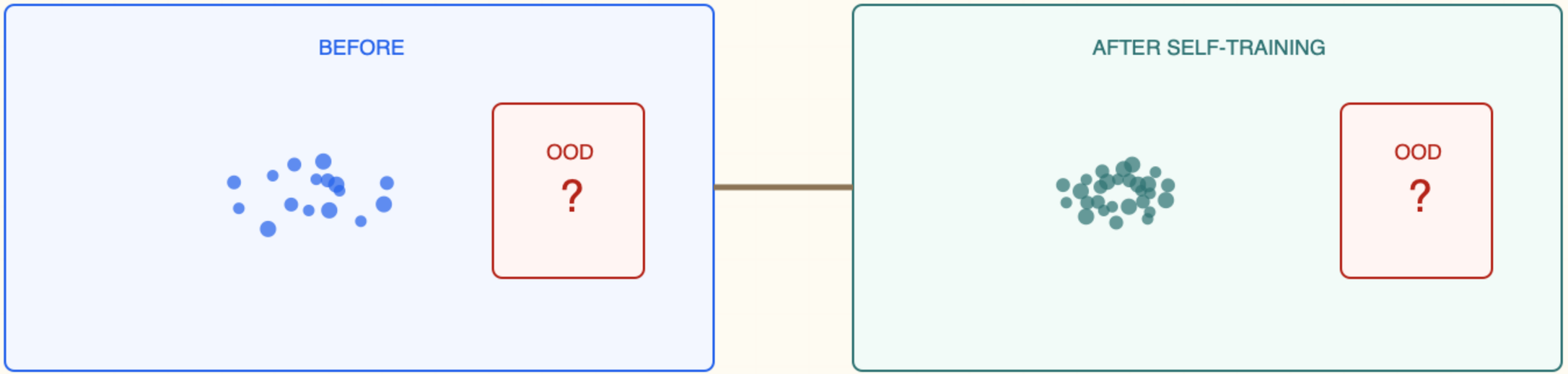


SUPPORT BOUNDARY

PART 6 37:34-40:53 / 63:12-67:58

Self-Training Amplifies Reachable Behavior

It can make supported strategies more reliable; it is not guaranteed to invent a genuinely absent or OOD strategy.



CAVEAT

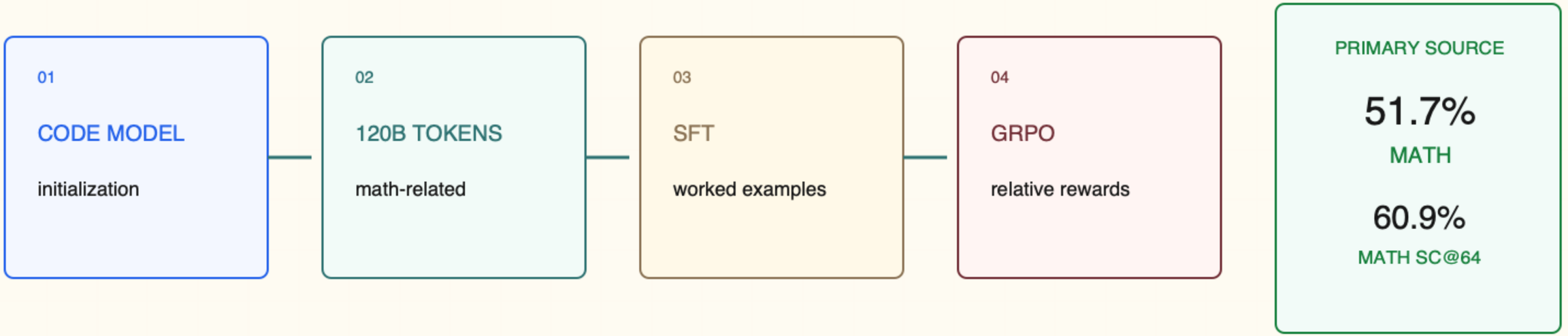
Lecture hypothesis, not a theorem: absent strategies may require new data, tools, curricula, or search.

DEEPSEEKMATH STACK

PART 6 40:53-46:25

DeepSeekMath Is a Full Training Stack

Code initialization, 120B math-related tokens, SFT, and GRPO jointly precede the reported result.



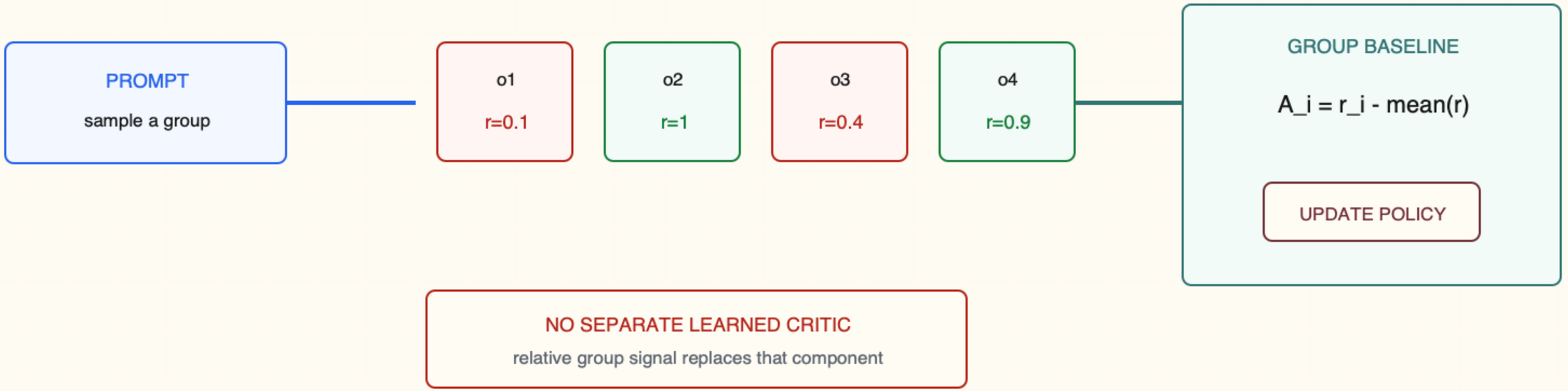
SOURCE CORRECTION: NOT AIME

RELATIVE GROUP FEEDBACK

PART 6 43:49-47:46

GRPO Uses Relative Group Feedback

Sample a group, normalize rewards within it, and update the policy without a separately learned critic.



SIGNAL GEOMETRY

PART 6 47:46-52:17 / 55:21-57:32

No Reward Variation Means No Relative Signal

All-wrong and all-correct groups collapse the within-group advantage; mixed groups carry the update signal.

CAVEAT

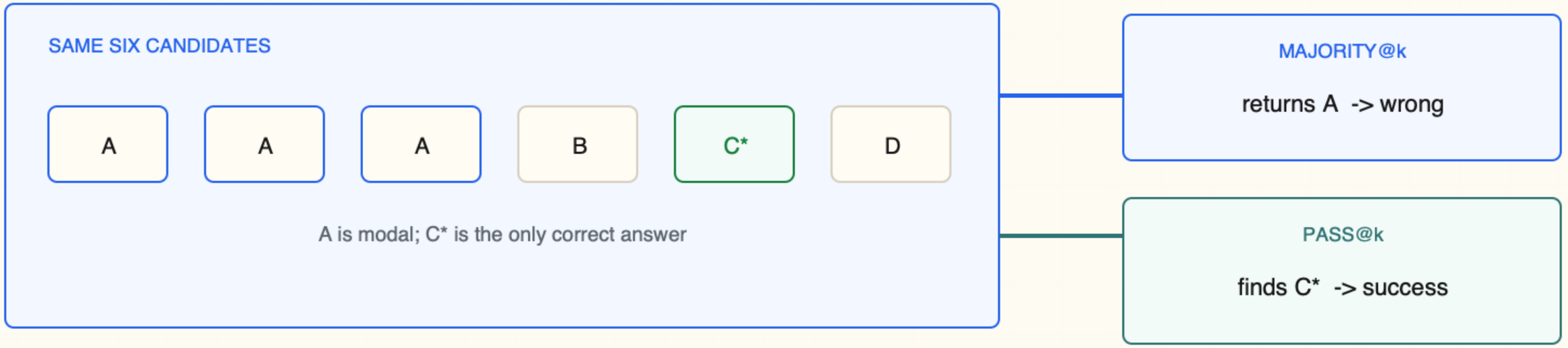
Dynamic sampling restores variation but also changes which prompts define the effective training distribution.

EVALUATION SEMANTICS

PART 6 52:17-53:05 / 63:12-65:11

Majority@k and Pass@k Measure Different Things

One rewards consensus around the modal answer; the other asks whether any sampled candidate succeeds.



CONSENSUS != COVERAGE

DAPO STABILIZATION

PART 6 53:05-61:49

DAPO Stabilizes RL With Four Coupled Controls

Clip-Higher, dynamic sampling, token-level loss, and overlong reward shaping form one cumulative recipe.

01

CLIP-HIGHER

protect exploration

02

DYNAMIC SAMPLE

keep reward variation

03

TOKEN LOSS

balance response lengths

04

OVERLONG SHAPE

bound truncation

PAPER SETUP

30 -> 50

AIME24 avg@32

Qwen2.5-32B

cumulative recipe

CAVEAT

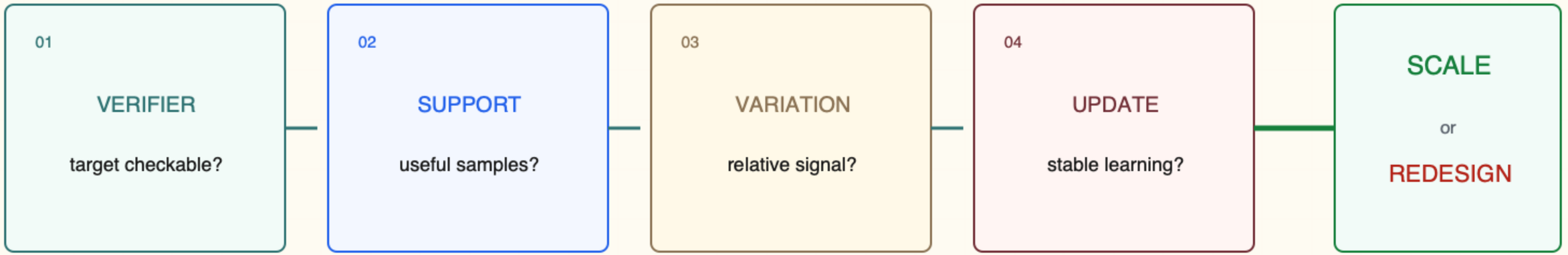
Setup-specific cumulative result; it is not a universal threshold or an isolated effect of one control.

METHOD DIAGNOSTIC

PART 6 61:49-72:34

Diagnose the Signal Before Scaling the Loop

Generated data helps only when the verifier, exploration support, reward variation, and update path are adequate.



A FAILED GATE IS AN ENGINEERING RESULT

Add supervision, widen exploration, repair rewards, or stop.