Paper discussion / AAAI 2026

MetaAct-RL

Training Language Models for Reasoning Through Meta-Action-Based Reinforcement Learning

Can a model learn when to continue, check, or revise its reasoning?

Zhiheng Xi et al. · Presented by Yash Saxena

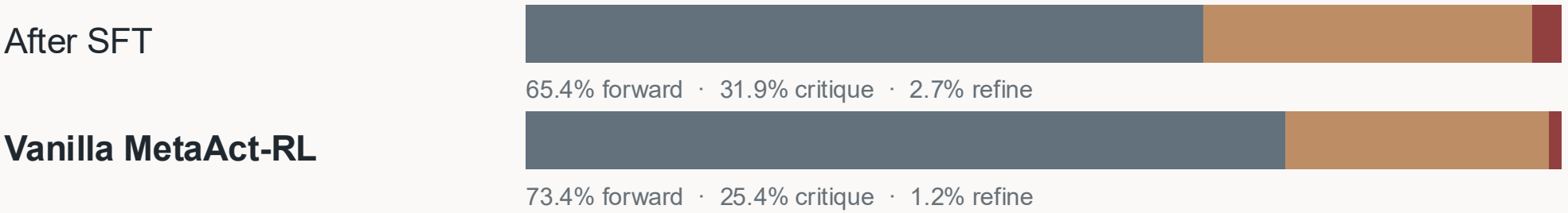
Why: outcome rewards can lose useful reasoning behavior

Ordinary CoT-RL

Question → Reasoning steps → **Answer** → **0 or 1**

A correct final answer says little about the intermediate steps.

Action proportions after meta-action SFT and vanilla RL



The shift toward forward reasoning motivated the full method.

What: can the model choose how to reason?

> Can outcome-based RL train a language model to reason, critique, and revise at the right moments?

Example from the paper: 8 oz of vitamin water has 125 calories. How many in 12 oz?

FORWARD REASONING

125 × 12 = 1,500 calories

→

CRITIQUE

That ignores the 8 oz starting amount.

→

REFINEMENT

(125 ÷ 8) × 12 = 187.5 calories

Each meta-action is generated text. The model selects its next behavior through normal token generation.

How: three training components

01

Meta-action SFT

Teach the model examples of reason → critique → refine

24,307 synthesized traces

02

Length-aware RL

Final-answer accuracy plus a clipped length bonus or penalty

KL regularization retained

03

Key-state restart

Practice continuing from a step where reasoning went wrong

50% key state, 50% question

RL still evaluates correctness at the final answer, not at each reasoning step.

1 · Supervised warm start

Actor
Draft a solution

→

Critique model
Inspect the reasoning

→

Actor
Revise the mistake

→

SFT
Learn the full trace

Example training trace

Question → **wrong step** → **critique** → **correction** → **correct answer**

24,307 examples

Source problems: MATH · GSM8K · AQuA

2 · Reward the final answer, then adjust for length

Baseline length per question

Average of 16 responses sampled from the SFT model

$$q = \text{clip}\left(\text{response length} \div \text{reference length},\ \mathbf{0.5},\ \mathbf{1.5}\right)$$

Correct final answer

$$\text{Reward} = 1 + q$$

Example: 1 + 1.3 = 2.3

Incorrect final answer

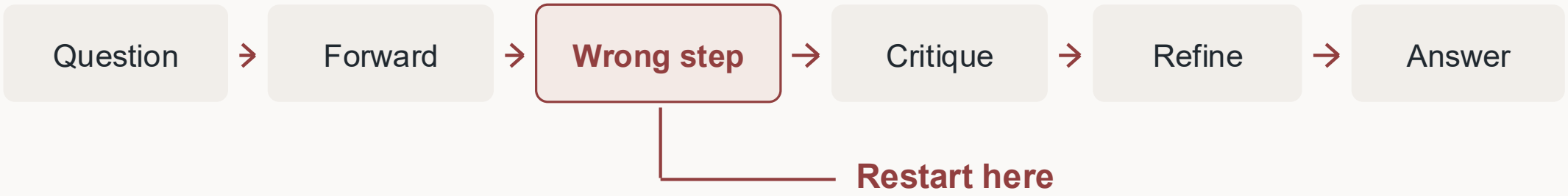
$$\text{Reward} = -0.5 \times q$$

Example: −0.5 × 1.3 = −0.65

The final-answer reward guides the full trajectory. No separate step-quality reward is used.

3 · Restart training after an incorrect step

One successful demonstration can contain a corrected mistake

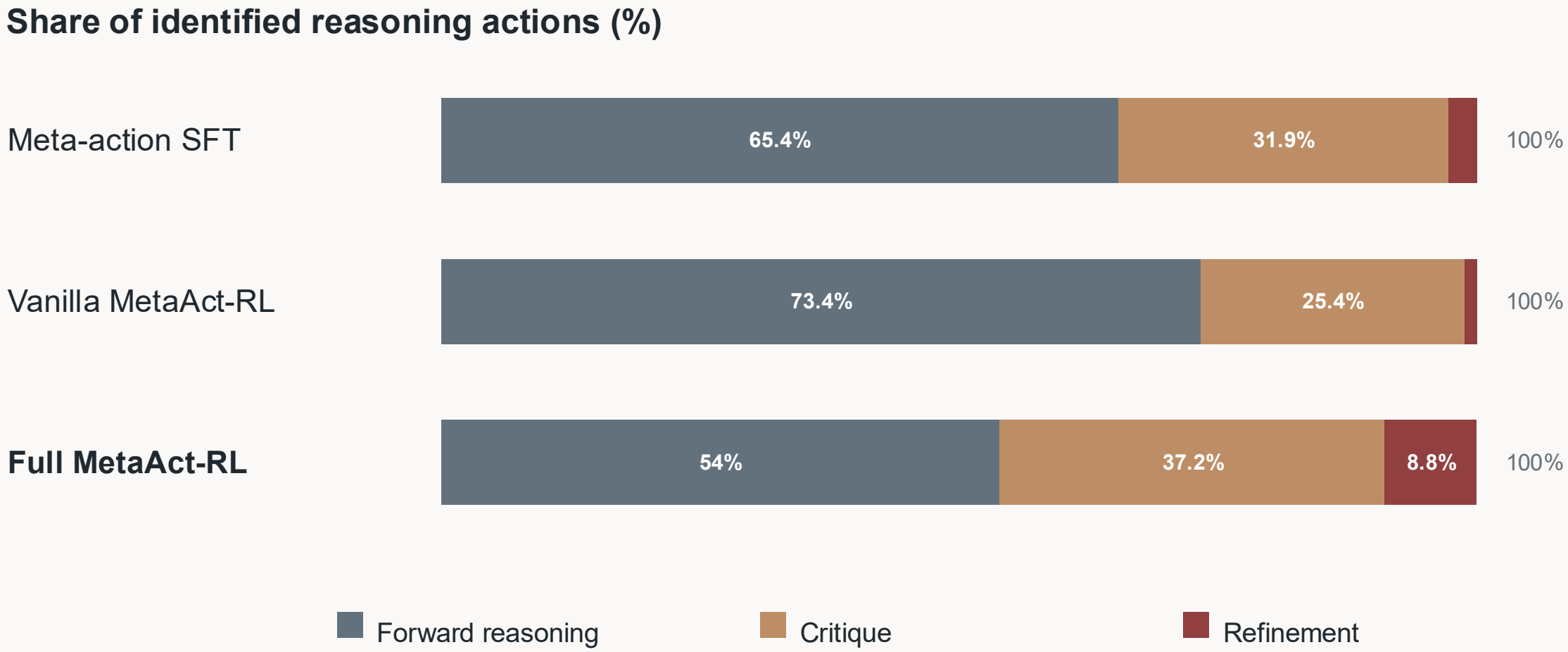


50% from the original question

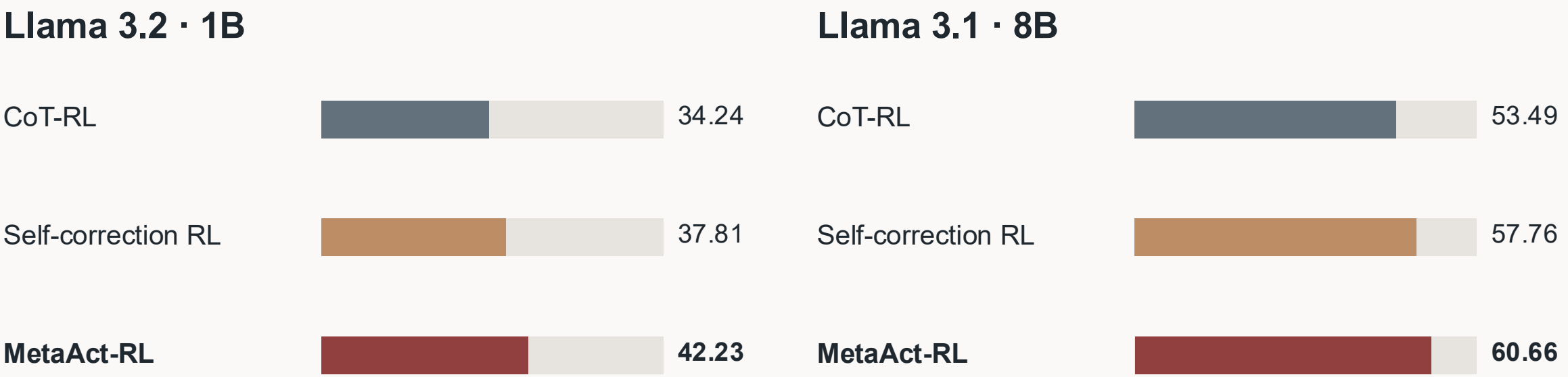
50% from a key-state prefix

The model generates a new continuation; reward is calculated from its final answer.

Does the method preserve a wider mix of actions?



Does it improve answer accuracy?



+7.99 pts over CoT-RL

+7.17 pts over CoT-RL

Qwen2.5-1.5B on AIME 2024: CoT-RL 28.33 → MetaAct-RL 35.83

Ablations: which pieces matter?

Training variant	AQuA	GSM8K	MATH	Mean length
Full MetaAct-RL	**48.43**	**52.92**	**25.00**	**513.3**
Without length reward	43.70	51.78	23.38	332.7
Without key-state restart	44.49	50.41	23.31	481.6
Restart from ordinary reasoning	43.48	51.25	23.03	476.8
Without upper length cap	**38.58**	**31.46**	**8.46**	**1,024.0**

Removing the upper cap: 513 → 1,024 tokens; MATH accuracy: 25.00 → 8.46.

Strengths, in the context of earlier work

Compared with STaR	Moves beyond fine-tuning on successful rationales: online RL trains continuations after intermediate errors.
Compared with SCoRe	Allows critique and refinement within a reasoning trajectory, rather than a fixed solution-then-correction round.
Compared with outcome-only CoT-RL	Shows why learning critique through SFT alone can fade during RL, then tests mechanisms designed to preserve it.
Evidence quality	Multiple model sizes, stronger self-correction baseline, component ablations, and a documented reward-hacking failure.

Weaknesses: where the evidence is incomplete

01 How are incorrect steps located?

The pool gains new RL trajectories with correct final answers, but the paper does not specify error localization for them.

02 Are critiques actually useful?

There is no direct per-step critique-quality reward; action proportions alone do not establish correct self-diagnosis.

03 Is the gain worth the extra compute?

No main comparison holds inference tokens or latency equal, despite a reward that favors longer correct traces.

04 Does it transfer outside math?

All reported evaluations are mathematical reasoning. Other NLP tasks remain untested.

A useful next test: evidence-grounded question answering

Candidate adaptation

Reason	Draft an answer from retrieved documents
Critique	Check each claim against cited evidence
Refine	Repair or remove unsupported claims

Key experiment Model-found vs. gold-located errors, at the same inference-token budget.

Takeaway: strong mathematical gains; error localization and step quality still need direct tests.

References

Xi et al. (2026)
MetaAct-RL: Training Language Models for Reasoning Through Meta-Action-Based Reinforcement Learning

AAAI 2026

Zelikman et al. (2022)
STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning

NeurIPS 2022

Kumar et al. (2025)
Training Language Models to Self-Correct via Reinforcement Learning (SCoRe)

ICLR 2025

Guo et al. (2025)
DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning

Nature 2025

Tyen et al. (2024)
LLMs cannot find reasoning errors, but can correct them given the error location

Findings ACL 2024

Huang et al. (2024)
Large Language Models Cannot Self-Correct Reasoning Yet

ICLR 2024

Paper: https://doi.org/10.1609/aaai.v40i40.40694