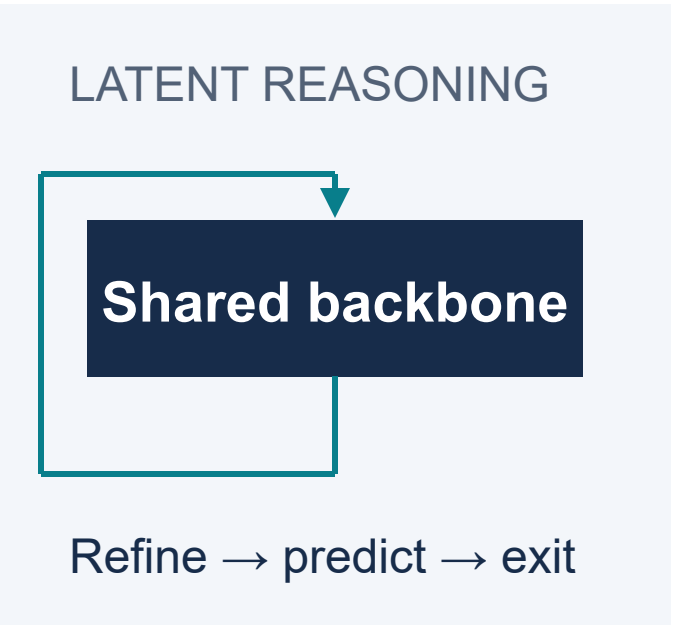
NEURIPS 2026

LoopRPT

Reinforcement Pre-Training for Looped Language Models



Stronger latent reasoning.

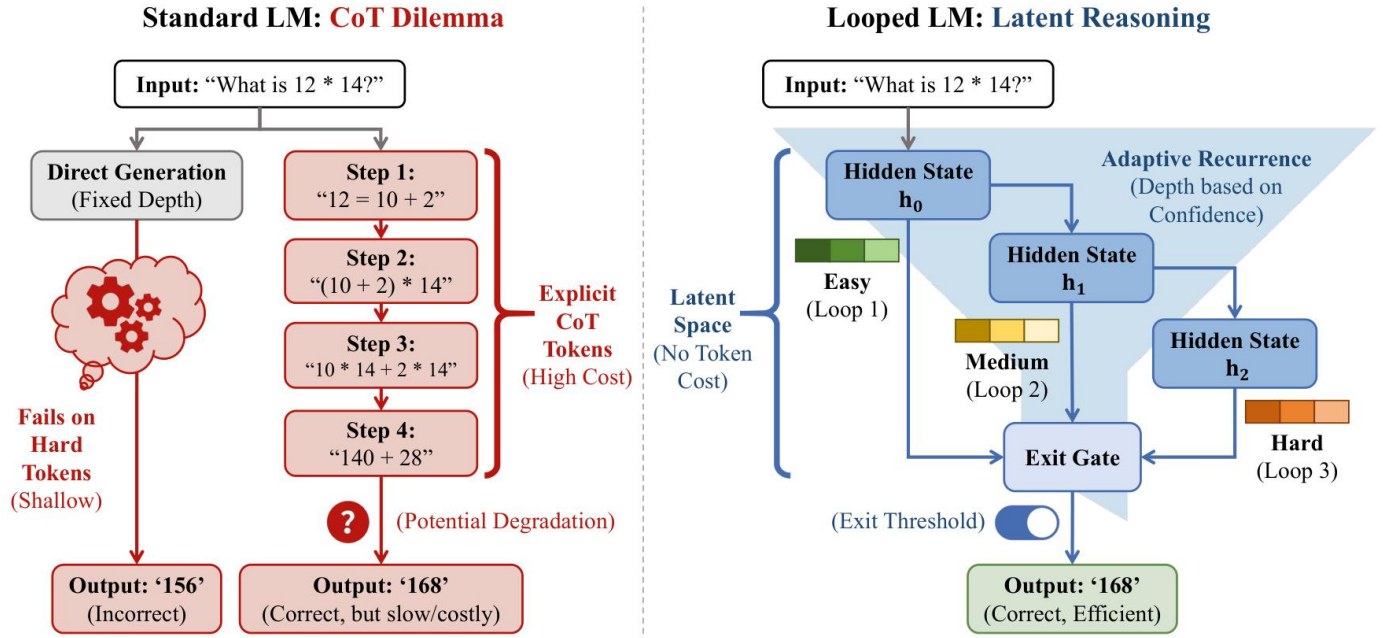
Fewer iterations.

Guo Tang* · Shixin Jiang* · Heng Chang* · Zihan Zhang · Nuo Chen
Yuhan Li · Huiming Fan · Jia Li · Ming Liu · Bing Qin

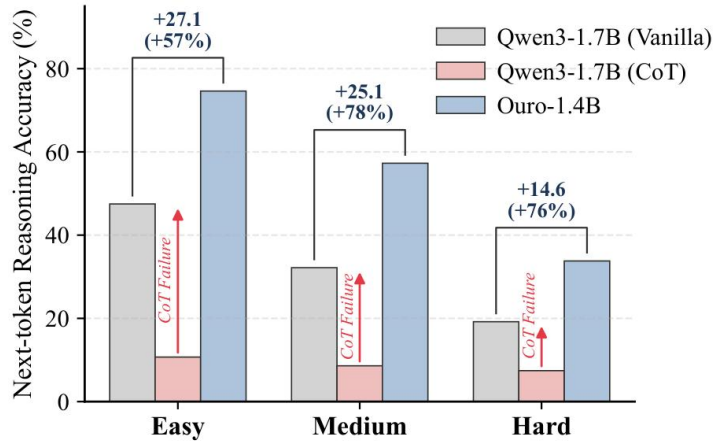
Harbin Institute of Technology · Peng Cheng Laboratory · Tsinghua University · Harvard University
Tencent HY Foundation Model Team · HKUST (Guangzhou) * Equal contribution

MOTIVATION

Why reinforce latent reasoning?



(a) Comparison of explicit CoT reasoning and latent space reasoning.

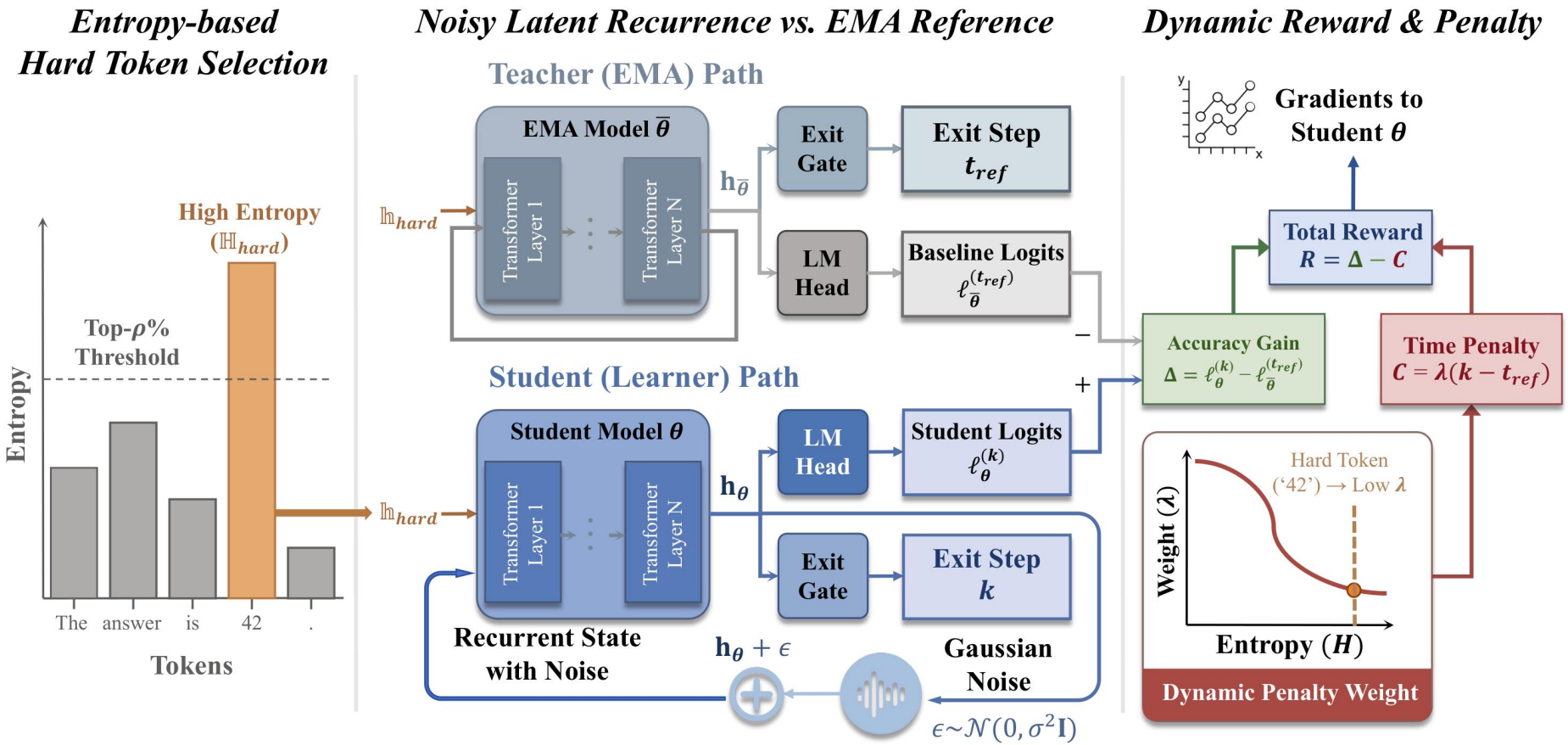


(b) Performance comparison between Ouro-1.4B and Qwen3-1.7B on next-token reasoning tasks across difficulty.

Latent steps → Next token → Output reward

Which latent iteration actually helped?

Focus learning on hard token positions



1 Select high-entropy tokens

2 EMA teacher reference

Reward useful computation; learn when to exit

$$R(k) = \left(\ell_{\theta}^{(k)} - \ell_{\bar{\theta}}^{(t_{\text{ref}})}\right) - \lambda_t(k - t_{\text{ref}})$$

predictive gain − difficulty-aware time penalty

Easier tokens: stronger compute penalty. Harder tokens: more room to reason.

Noisy latent rollouts → **Exit policy**

Gaussian exploration — Group-relative advantages

Noise-free forward pass → **Intermediate states**

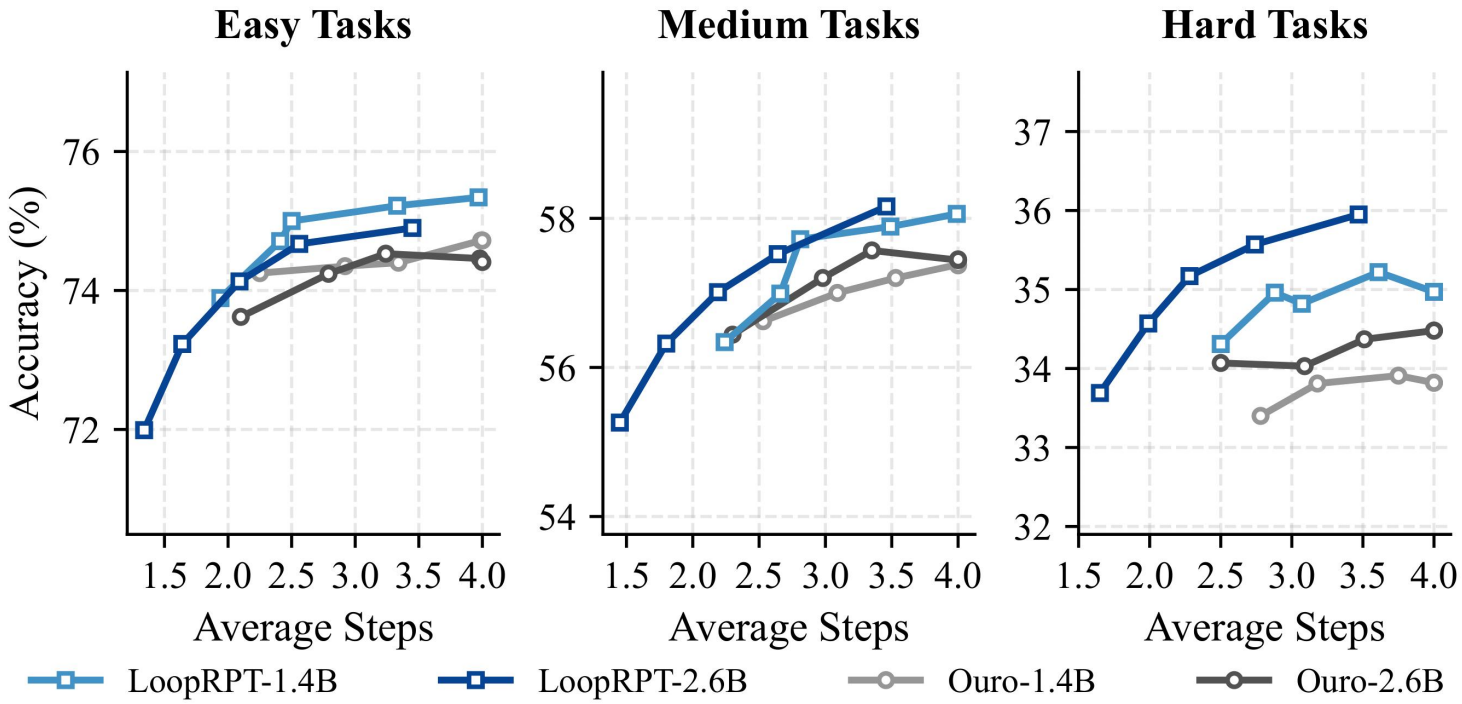
Reward-shaped step weights — Step-weighted token prediction

Entropy and teacher KL regularize the joint objective.

ACCURACY AND COMPUTE

A better accuracy–compute trade-off

Ouro 1.4B / 2.6B · Omni-Math: 200 held out · maximum 4 latent steps



2.6B adaptive Hard

+2.89 pp

34.35% → 37.24%

−35.0%

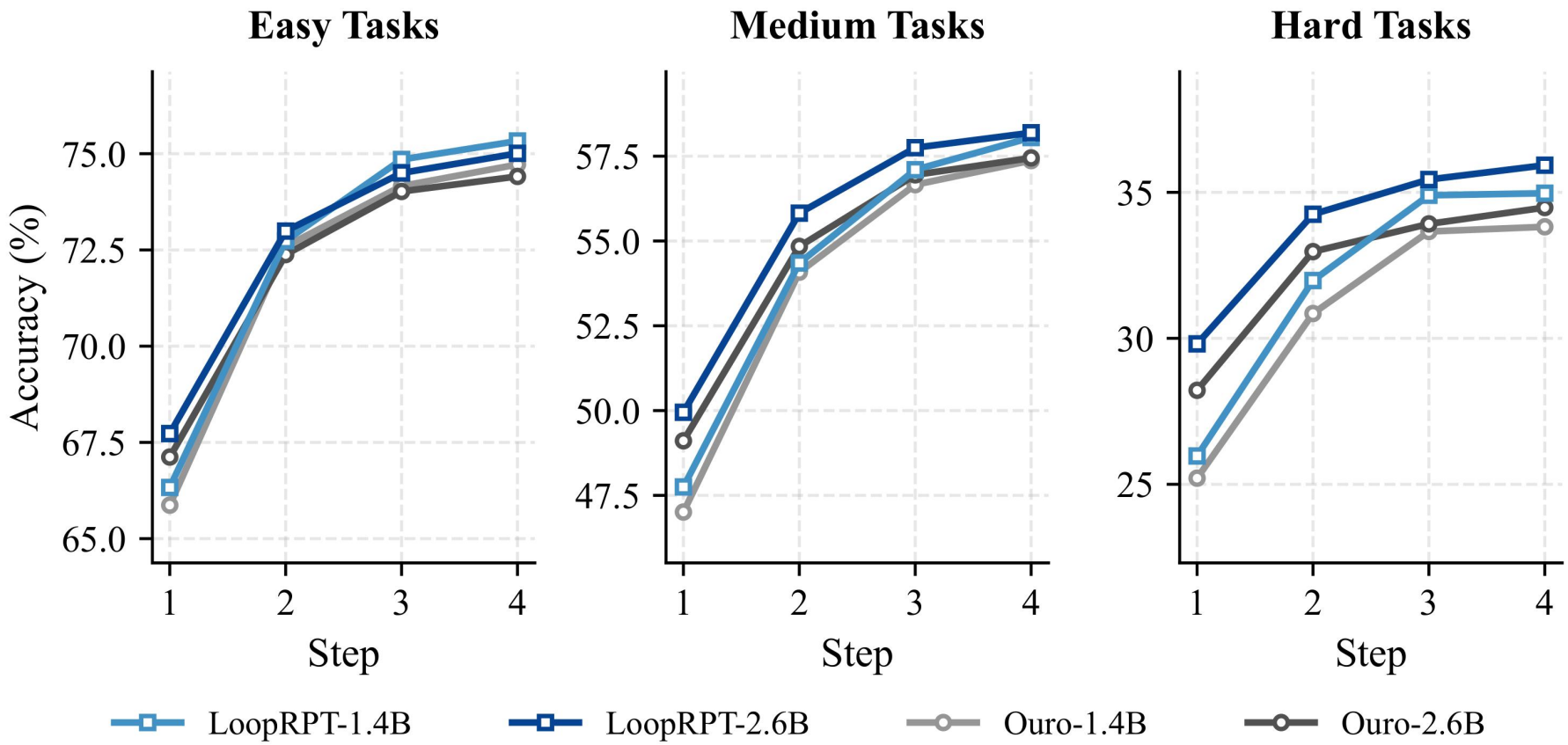
average latent steps

3.51 → 2.28

NTR: correct next-token accuracy given context.

Hard: token-entropy subset. Average steps are a compute proxy.

Better early states explain the improvement



Early-step predictions improve, beyond changing exit decisions.

LoopRPT · Fig. 7 · Intermediate-state readouts; forced-depth evaluation is a separate experiment

OBJECTIVE ATTRIBUTION · OURO-2.6B

The objective matters beyond extra training

Objective	NTR-Hard (%) Adaptive exit	GSM8K (%) 3-shot CoT
Base checkpoint	34.35	81.76
Continued-NTP	35.21	82.41
GRPO / RLVR	33.00	6.63
LoopRPT	**37.24**	**85.36**

Same checkpoint, data, step budget and schedule.

Training arms differ in objective; GRPO failure is specific to this configuration.

DOWNSTREAM TRANSFER

Transfer is strongest on math and code here

Benchmark	Scale	Ouro	LoopRPT	Δ (pp)
MBPP+	1.4B	60.85	**63.76**	**+2.91**
GSM8K	2.6B	81.76	**85.36**	**+3.60**
HumanEval+	2.6B	70.13	**71.95**	**+1.82**
MMLU	2.6B	73.85	**73.91**	+0.06

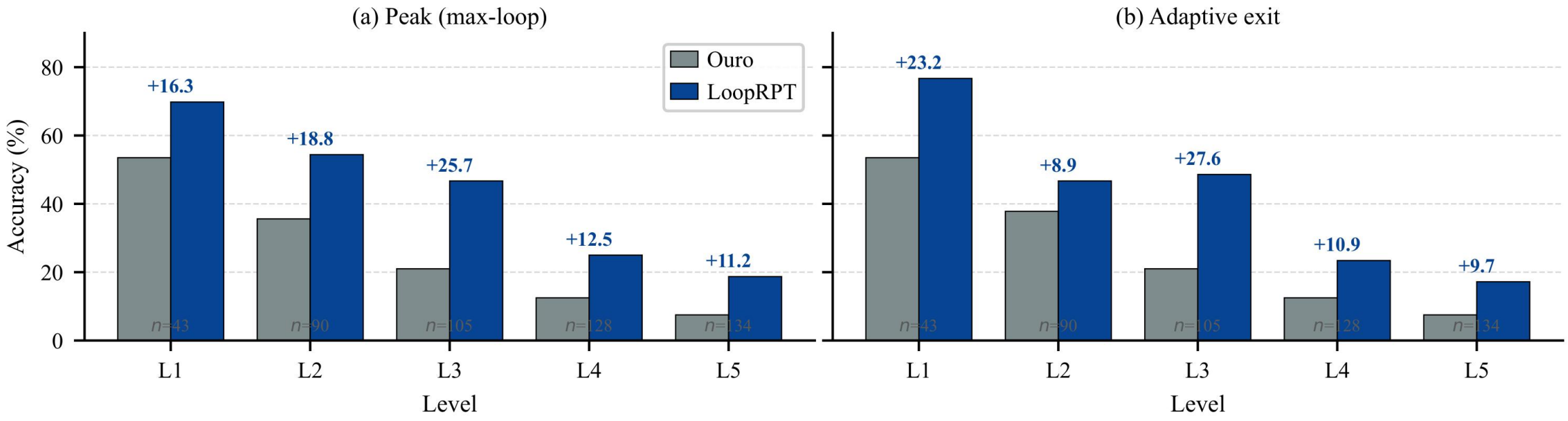
GSM8K: 3-shot CoT · Coding: pass@1 · MMLU: 5-shot logprobs

Transfer is task-dependent; general-task gains can be small.

LoopRPT · Tables 2 and 12, selected rows · Protocols: Table 7 · All values are percentages

HARD MATHEMATICS

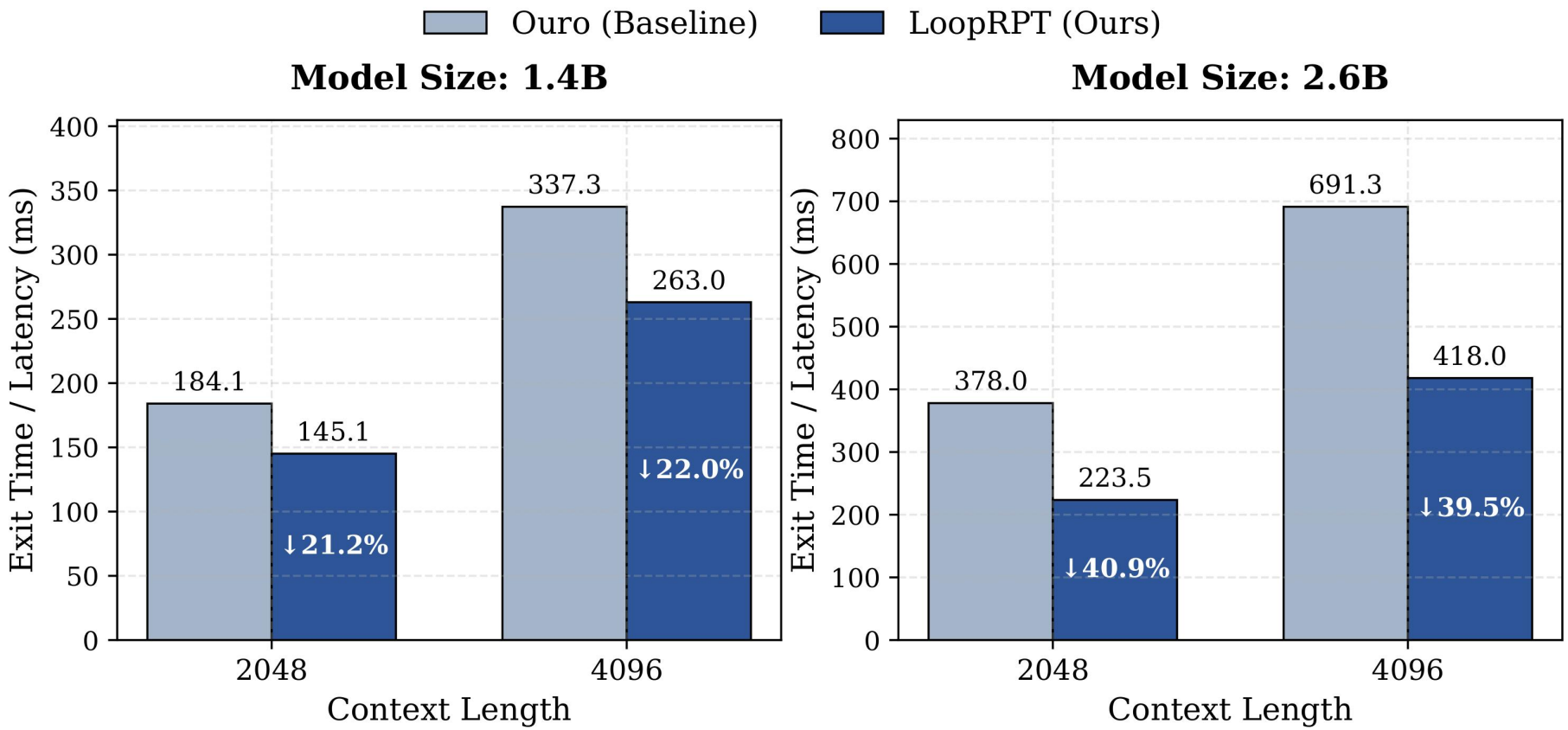
The gain extends to hard mathematics



Ouro-1.4B improves across all five MATH500 levels.

PHYSICAL EFFICIENCY

Fewer iterations, lower measured latency



2.6B · 4,096-token context · one A100: 39.5% lower latency.

Reinforce the computation inside the loop

Hard tokens

Step-wise rewards

State + exit learning

Improve accuracy and computation together.

By rewarding latent steps · Demonstrated on Ouro 1.4B / 2.6B

Thank you.