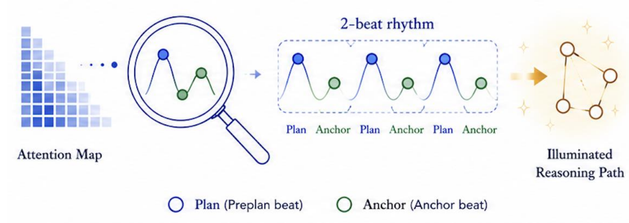
Attention *Illuminates* LLM Reasoning

The Uncovered Preplan-and-Anchor Rhythm Enables Fine-Grained Policy Optimization

Yang Li, Zhichen Dong, Yuhan Sun, Weixun Wang, Shaopan Xiong, Yijia Luo, Jiashun Liu, Han Lu, Jiamang Wang, Wenbo Su, Bo Zheng, Junchi Yan
Shanghai Jiao Tong University · Alibaba ATH Group · Shanghai Innovation Institute



1. Key Insight: The Preplan-and-Anchor Rhythm

- Not all tokens reason equally.
- Some prepare the next step.
- Some anchor the future path.

Together: a repeating two-beat rhythm

A. Local Head Attention: Chunk Builders

B. Global Head Attention: Anchor Trackers

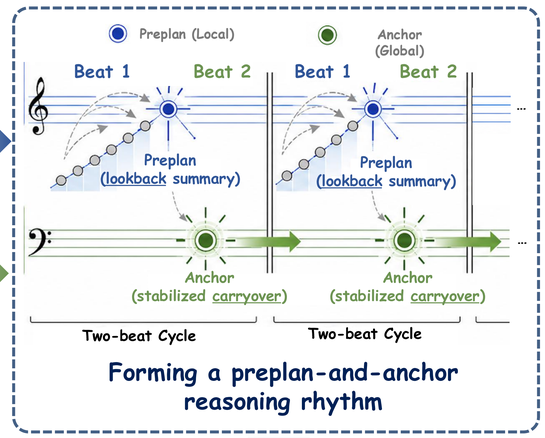
$$D_{head} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{i-1} A_{i,j}\cdot(i-j)$$

C. Coupled Rhythm Signals

Local Chunks: sawtooth lookback → preplan

This model finally finds more hidden patterns

Global Anchors: future attention stripes → anchor



Forming a preplan-and-anchor reasoning rhythm

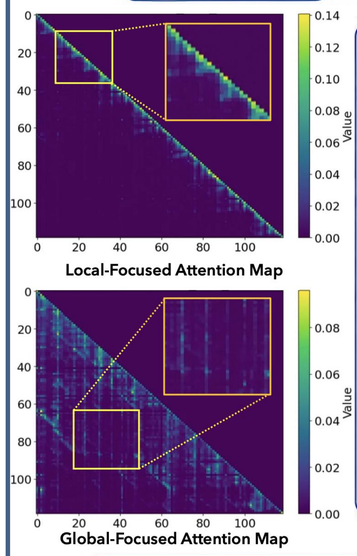
Fine-Grained RL Credit Assignment

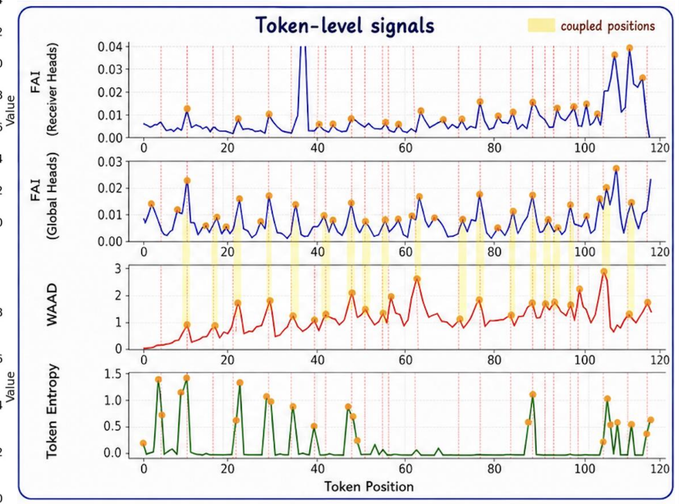
Attention rhythm → token credit modulation

$A_t \rightarrow \gamma_t A_t$

2. How it is Measured: Attention Maps & Token Metrics

Split heads (local / global) → Compute token metrics (WAAD + FAI) → Detect rhythm pairs (WAAD peak → FAI peak)





WAAD — How far does token t look back? → preplan / chunk onset

FAI — How much do future tokens rely on token t? → anchor / semantic pivot

Coupling — FAI peak follows or coincides with WAAD peak → preplan-anchor rhythm

3. Rhythm-Aware RL: Credit Assignment Strategies

1 Uniform Credit → 2 Detect Critical Rhythm Tokens → 3 Amplified Token Credit

Same sequence reward for every token

Key reasoning steps are diluted.

Preplan Preplan Anchor

A Local-Chunk Credit — WAAD peak / chunk onset

B Global-Anchor Credit — High FAI / downstream influence

C Coupled Rhythm Credit — Preplan → Anchor

Attention signals (WAAD peaks & FAI) select important rhythm Tokens.

Scaled Advantage: $A_t \rightarrow \gamma_t A_t$

$\gamma_t > 1$ only for rhythm-selected tokens.

- γ_t^{local}: More Credit for Introductory Tokens of Semantic Groups.
- γ_t^{global}: More Credit for Anchor Tokens for Subsequent Reasoning.
- $\gamma_t^{combined}$: More Credit for Reallocated Semantic Influence of Semantic Groups.

4. Experiment Results

Rhythm-Aware RL outperforms baselines on 1K and 8K math reasoning tasks.

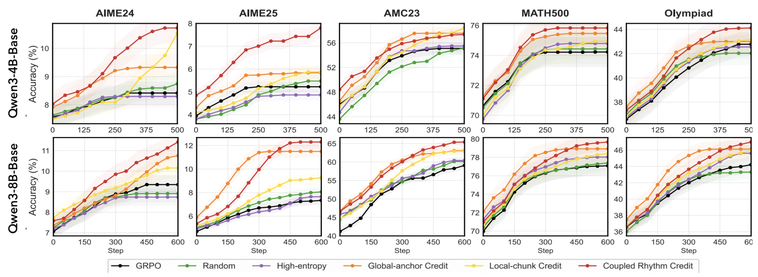
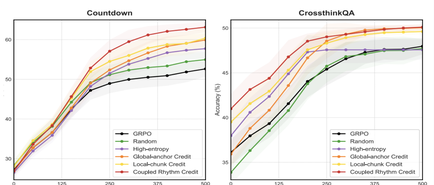


Table 4. Results of math reasoning. AIME 24&25: avg@16; others: pass@1.

Method	AIME24	AIME25	AMC23	MATH	Olympiad	Avg.
Qwen3-4B-Base with 1K Length						
GRPO	8.4	5.2	55.1	74.2	42.8	37.1
+ random credit	8.7	5.5	55.2	74.4	42.0	37.1
+ high-entropy credit	8.3	4.9	55.5	74.8	42.5	37.2
+ global-anchor credit	9.3 +0.9	5.8 +0.6	57.6 +2.5	75.5 +1.2	43.0 +0.2	38.2 +1.1
+ local-chunk credit	10.5 +2.1	5.9 +0.7	**58.4 +3.3**	74.9 +0.7	43.1 +0.4	38.6 +1.5
+ coupled rhythm credit	**10.7 +2.3**	**7.8 +2.6**	57.4 +2.3	**75.8 +1.6**	**44.1 +1.3**	**39.2 +2.1**
Qwen3-8B-Base with 1K Length						
GRPO	9.3	7.3	59.1	77.1	44.2	39.4
+ random credit	8.9	8.1	60.1	77.4	43.3	39.6
+ high-entropy credit	8.7	7.7	60.4	78.0	45.6	40.1
+ global-anchor credit	10.8 +1.4	11.5 +4.2	63.1 +4.0	78.9 +1.8	46.1 +1.9	42.1 +2.7
+ local-chunk credit	10.1 +0.8	9.3 +1.9	62.9 +3.8	78.4 +1.3	45.9 +1.7	41.3 +1.9
+ coupled rhythm credit	**11.4 +2.1**	**12.3 +5.0**	**65.4 +6.3**	**79.7 +2.6**	**47.0 +2.8**	**43.2 +3.8**
Qwen3-8B-Base with 8K Length						
GRPO	19.5	16.1	57.6	81.0	49.9	44.8
+ random credit	18.0	16.6	57.1	82.0	50.0	44.7
+ high-entropy credit	19.3	15.4	57.8	81.2	48.6	44.5
+ global-anchor credit	21.7 +2.3	19.5 +3.4	**61.1 +3.5**	82.1 +1.2	51.2 +1.3	47.1 +2.3
+ local-chunk credit	**22.4 +2.9**	**20.4 +4.3**	59.2 +1.6	82.5 +1.5	51.5 +1.6	47.2 +2.4
+ coupled rhythm credit	22.0 +2.5	19.9 +3.8	59.3 +1.7	**82.9 +1.9**	**52.2 +2.3**	**47.3 +2.5**

Rhythm-Aware RL proves effective on diverse reasoning task types.



Our method generalizes across model families and scales.

Table 6. Robustness across model families and scales. Each benchmark entry reports the result of coupled rhythm credit, with the subscri denoting the absolute improvement over GRPO under the same setting. All experiments use a 1K response length.

Category	Model	AMC23	MATH500	Olympiad	Minerva	GSM8K	Avg.	Rel. Gain
Other families	Llama3.2-3B	$4.8_{+1.0}$	$6.3_{+0.9}$	$2.9_{+0.7}$	$3.3_{+0.1}$	$11.0_{+1.5}$	$5.7_{+0.9}$	+18.8%
	Llama3.1-8B	$5.2_{+2.0}$	$7.7_{-0.4}$	$3.9_{+0.7}$	$6.0_{+0.7}$	$20.8_{+2.1}$	$8.7_{+1.0}$	+13.0%
	DeepSeek-R1-Distill-Llama-8B	$45.6_{+3.2}$	$64.2_{+1.4}$	$31.9_{+3.0}$	$25.8_{+1.5}$	$79.2_{+1.2}$	$49.34_{+2.06}$	+4.36%
Other scales	Qwen3-0.6B	$11.2_{+2.9}$	$37.9_{+7.4}$	$10.1_{+2.5}$	$12.7_{+5.4}$	$59.7_{+10.7}$	$26.32_{+5.78}$	+28.14%
	Qwen3-1.7B	$20.7_{+4.8}$	$55.9_{+3.8}$	$18.5_{+0.7}$	$24.2_{+6.1}$	$77.8_{+0.7}$	$39.42_{+3.22}$	+8.90%

The hyperparameters proves to be robust.

(b) WAAD/FAI hyperparameters

Sweep	Best / chosen Avg.	Avg. range
WAAD W	**38.6** at 10	37.0–38.6
FAI H_{lo}	**38.2** at 10	34.8–38.2
FAI H_{hi}	**38.2** at 100	37.4–38.2
GRPO baseline	37.1	–

Our method brings minimal time overhead.

Table 8. Computational overhead of credit assignment.

Model	Seq. Length	Avg. Step Time	Attn. Credit	Relative
Qwen3-4B	1,024	72 s (1.2 min)	1.6 s	2.2%
Qwen3-4B	8,192	300 s (5 min)	10.0 s	3.3%
Qwen3-8B	1,024	87 s (1.45 min)	1.7 s	2.0%