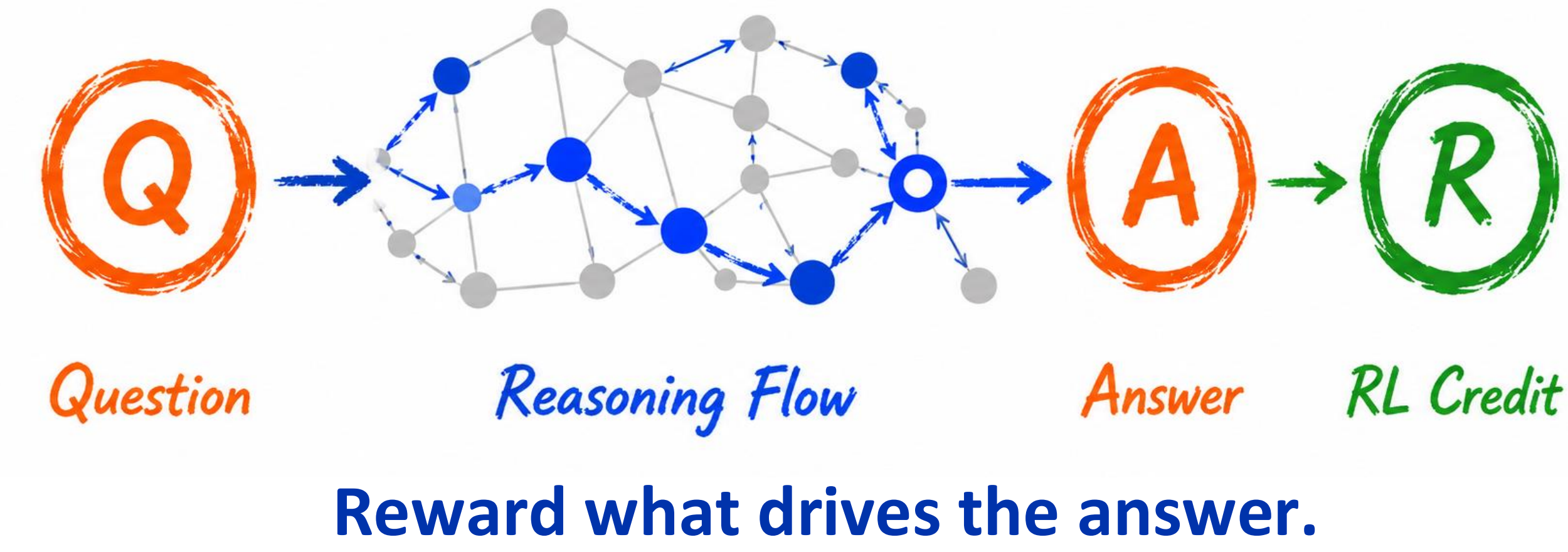


How does Reasoning Flow?

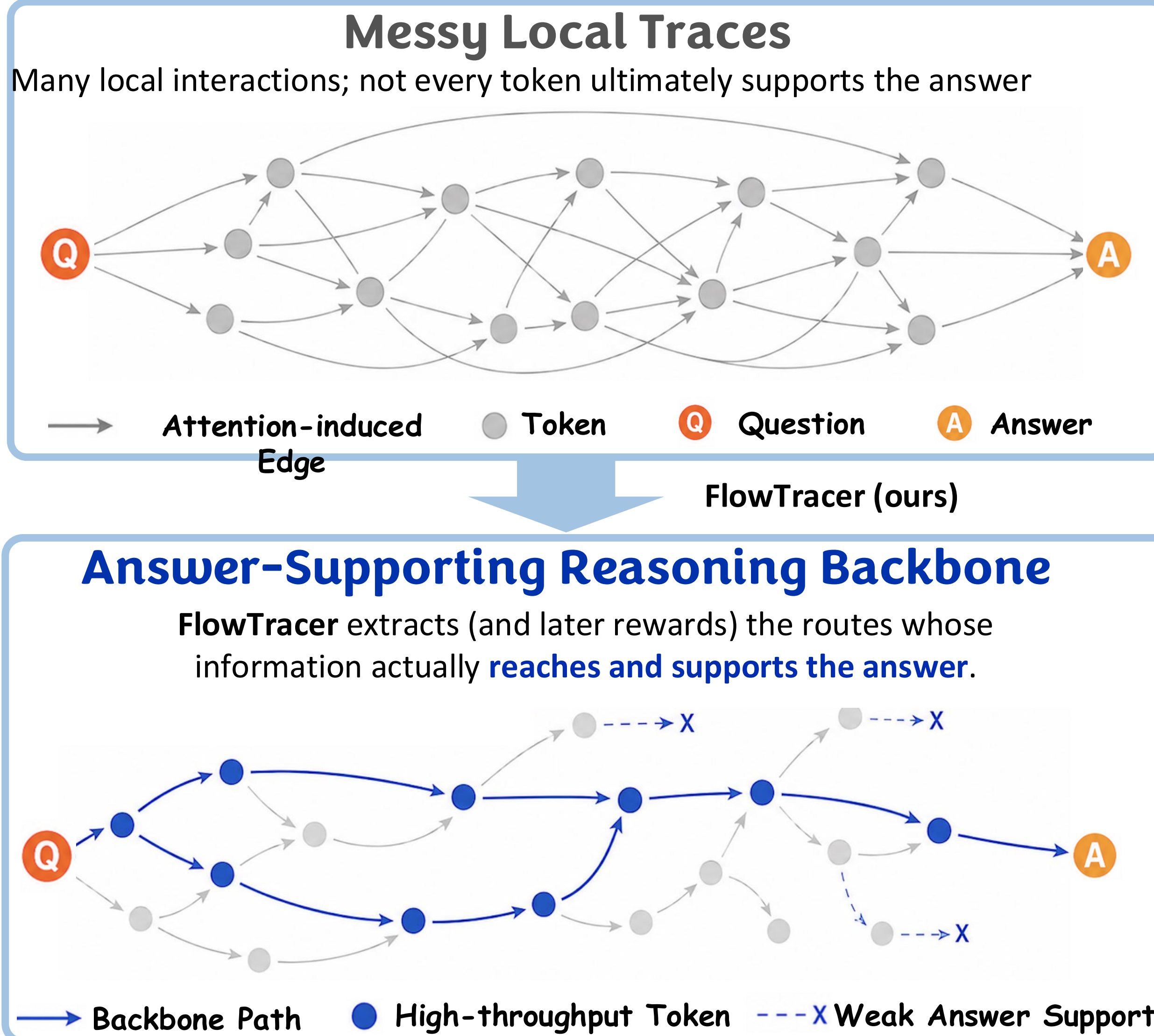
Tracing Attention-Induced Information Flow for Targeted RL in LLMs

Zhichen Dong*, Yang Li*, Yuhao Sun, Weixun Wang, Yijia Luo, Zinian Peng, Wenbo Su, Cheng Yu, Bo Zheng, Junchi Yan
Shanghai Jiao Tong University · Alibaba ATH Group · Shanghai Innovation Institute



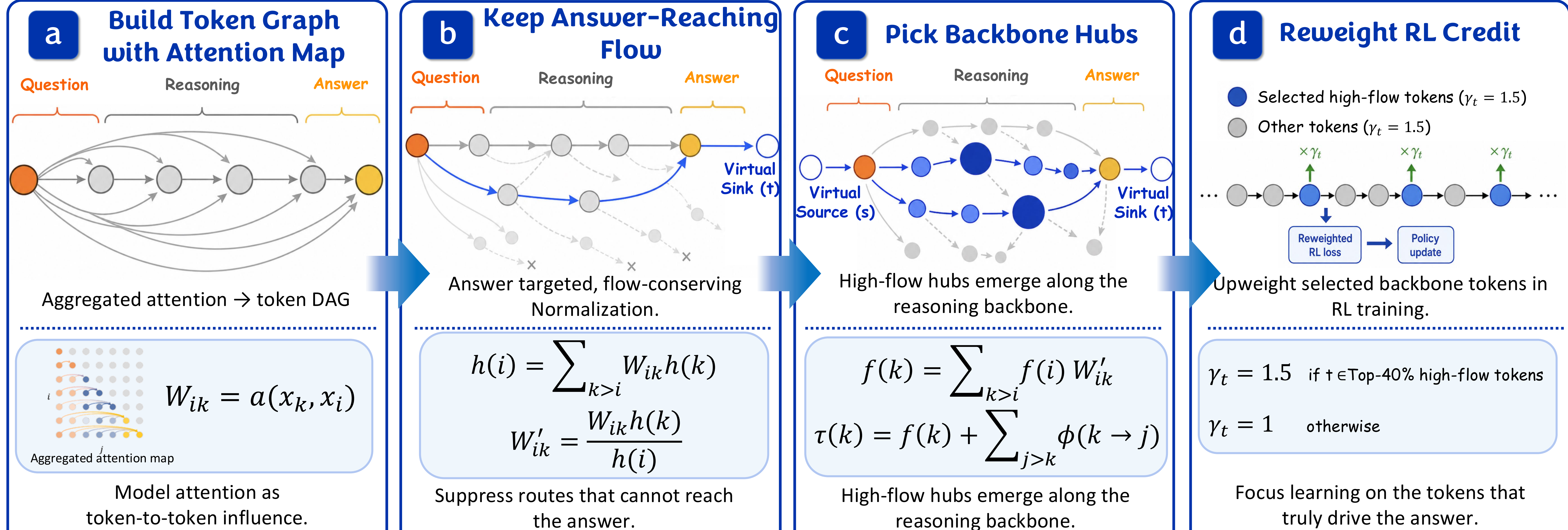
1 Core Intuition

Reasoning forms a token-level flow network from **Question** to **Answer**. **FlowTracer** isolates the **answer-supporting backbone** and assigns targeted RL credit to it.



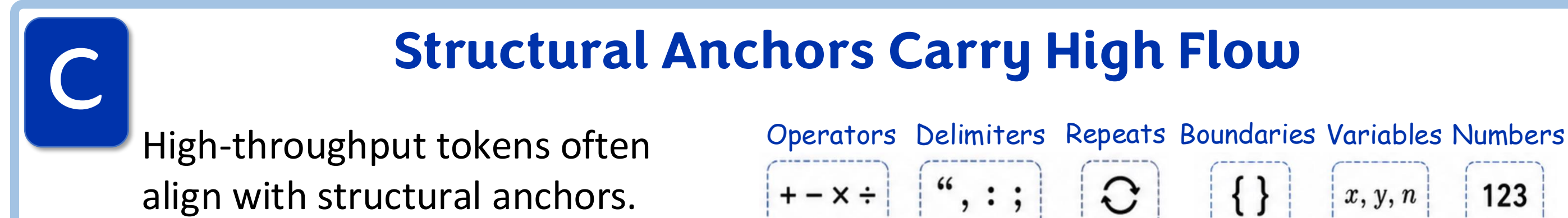
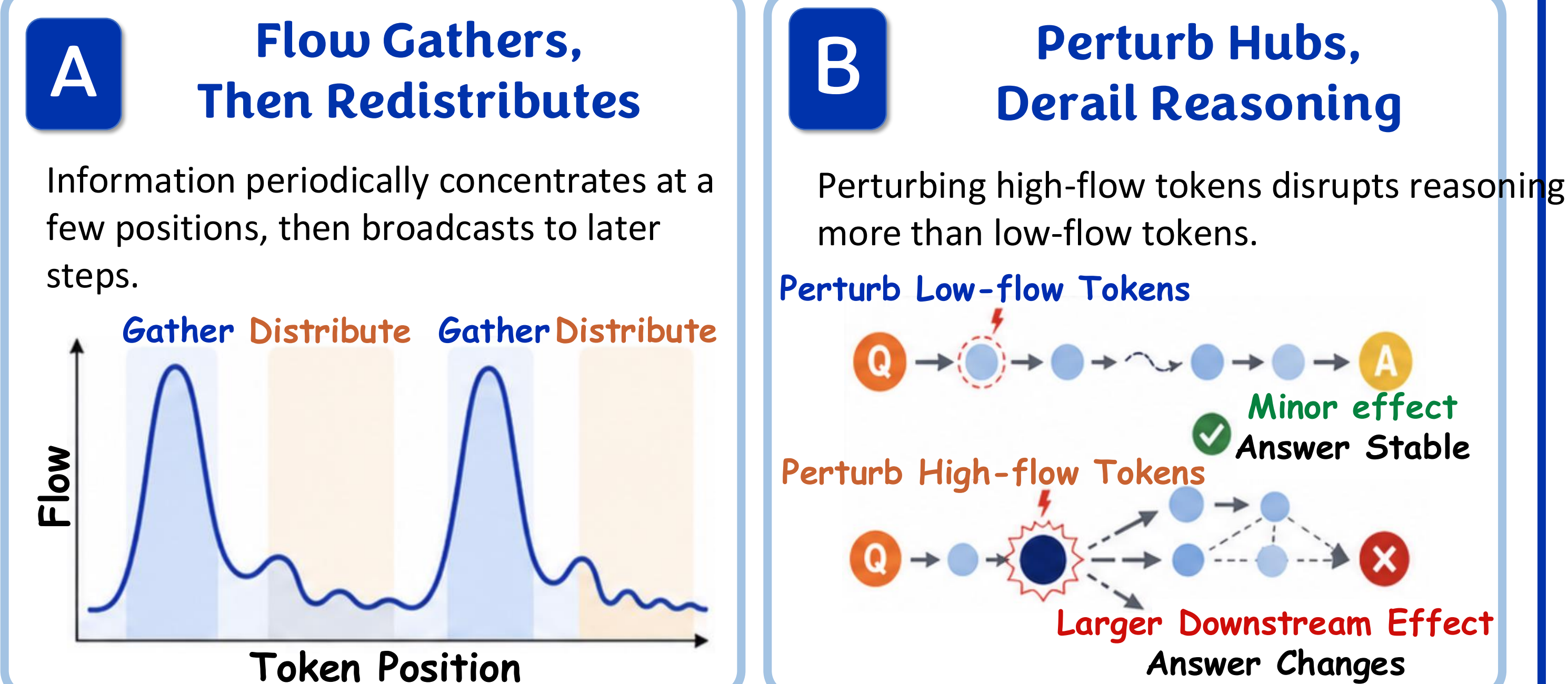
- ✓ View reasoning as an **information-flow network** across tokens.
- ✓ **Filter out** dead-end branches that never support the final answer.
- ✓ Concentrate **reward** on the backbone that truly drives the answer.

2 How We Trace the Reasoning Flow



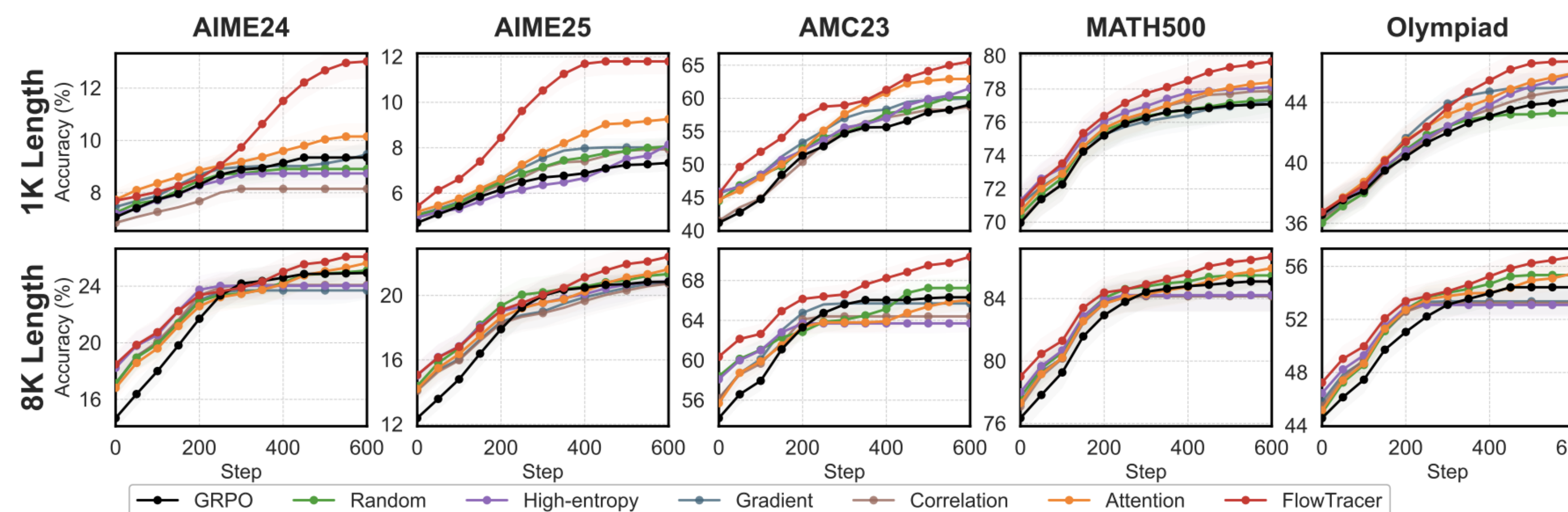
Attention-induced Graph → Answer targeted, Flow-conserving Normalization → Backbone Tokens → RL Credit Reweighting

3 What Reasoning Flow Reveals

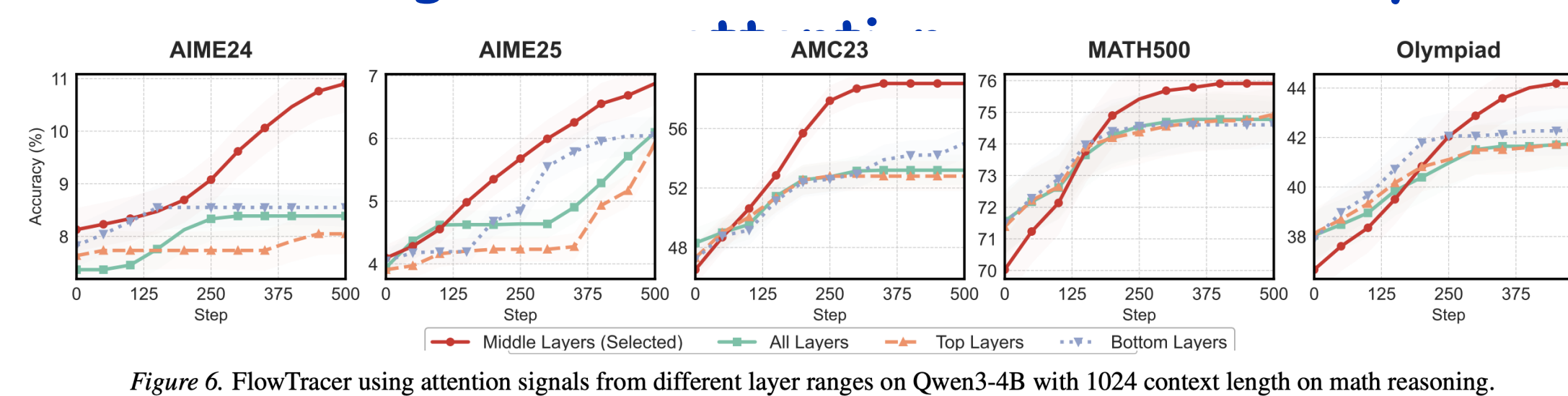


4 Experiment Results

FlowTracer maintains the **strongest training dynamics** on math reasoning benchmarks.



Reasoning flow concentrates in **middle-layer**



FlowTracer is applicable across **diverse reasoning paradigms**.

Method	Countdown	CrossThinkQA
GRPO	52.6	48.0
Random	55.0	47.6
High-entropy	57.7	47.6
Attention	60.4	49.6
FlowTracer	63.2 +10.6	50.2 +2.2

Low overhead: FlowTracer introduces only a marginal time overhead of 2.2%-4.5%

Table 6. Computational Overhead Analysis: Average training step time and FlowTracer's credit assignment time across different model sizes and context lengths.

Model	Context	Step Time	Flow Credit	Overhead
Qwen3-4B	1K	71.0 s	1.57 s	2.2%
	8K	189.7 s	8.55 s	4.5%
Qwen3-8B	1K	79.0 s	1.66 s	2.1%
	8K	275.3 s	11.04 s	4.0%

Table 2. Results of math reasoning on Qwen3-Base models. We compare performance between 1K and 8K context lengths across various methods. **Bold** denotes the best results.

Method	AIME24		AIME25		AMC23		MATH500		Olympiad		Avg.	
	1K	8K	1K	8K	1K	8K	1K	8K	1K	8K	1K	8K
Qwen3-4B-Base												
GRPO	8.4	19.5	5.2	16.1	55.1	57.6	74.2	81.0	42.8	49.9	37.1	44.8
Random	8.7	18.0	5.5	16.6	55.2	57.1	74.4	82.0	42.0	50.0	37.2	44.7
High-entropy	8.3	19.3	4.9	15.4	55.8	57.8	74.8	81.2	42.8	48.6	37.3	44.5
Gradient	8.6	20.3	4.2	19.6	53.2	57.8	74.1	82.3	43.3	51.3	36.7	46.3
Correlation	8.7	19.3	5.5	15.4	55.2	57.8	74.4	81.2	42.0	48.6	37.2	44.5
Attention	10.5	22.4	5.9	20.4	58.4	59.3	74.9	82.3	43.1	51.5	38.6	47.2
FlowTracer	10.9 +2.5	22.7 +3.2	6.9 +1.7	21.9 +5.8	59.0 +3.9	62.4 +4.8	75.9 +1.7	83.1 +2.1	44.2 +1.4	53.0 +3.1	39.4 +2.2	48.6 +3.8
Qwen3-8B-Base												
GRPO	9.3	24.9	7.3	20.8	59.1	66.3	77.1	85.1	44.2	54.4	39.4	50.3
Random	8.9	25.1	8.1	21.3	60.1	67.2	77.4	85.5	43.3	53.3	39.6	50.9
High-entropy	9.7	24.0	8.1	20.9	61.5	63.7	78.1	84.2	45.9	53.1	40.5	49.2
Gradient	9.5	23.7	8.0	20.8	60.0	65.7	77.3	84.2	45.0	53.4	40.0	49.5
Correlation	8.2	24.1	7.9	20.7	58.8	64.4	77.9	84.2	44.8	53.2	39.5	49.3
Attention	10.1	25.6	9.3	21.6	62.9	66.1	78.4	85.9	45.9	55.5	41.3	50.9
FlowTracer	13.0 +3.7	26.1 +1.2	11.8 +4.5	22.4 +1.6	65.6 +6.5	70.4 +4.1	79.7 +2.6	86.7 +1.6	46.7 +2.5	56.7 +2.3	43.4 +4.0	52.5 +2.1