

# ALPHAONE

## Reasoning Models Thinking Slow and Fast at Test Time



Junyu Zhang\*



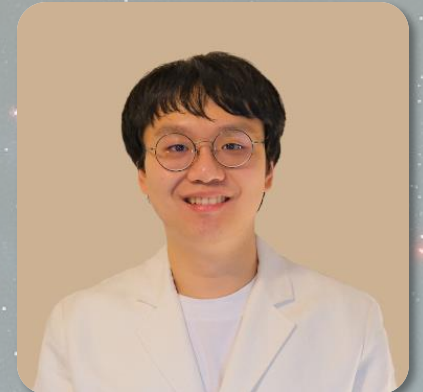
Runpei Dong\*



Jitendra Malik



Saurabh Gupta



Huan Zhang

**Other Authors:** Han Wang, Xuying Ning, Haoran Geng, Peihao Li, Xialin He, Yutong Bai

**NVIDIA, Aug 15, 2025**



# Background: Large Reasoning Models

Notice anything?

Thought for 28s

Yep—couple of red flags:

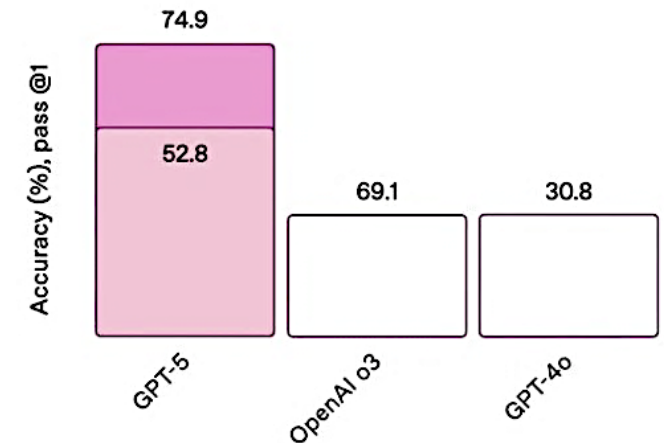
- **Apples-to-apples mismatch.** Only GPT-5 shows *both* “without” (52.8) and “with thinking” (74.9). o3 (69.1) and GPT-4o (30.8) are single numbers—unclear which condition they correspond to. The stacked vs hollow bars make it look like missing data.
- **Legend doesn’t map to the visuals.** Two purple dots in the legend, but o3/4o bars are just outlines. Viewers can’t tell which shade = which condition.
- **Labeling clarity.** Y-axis reads “Accuracy (%), pass@1”; flip to “Pass@1 accuracy (%)” and add ticks (0–100). For GPT-5 you label two numbers, others one—be consistent.
- **Naming consistency.** “OpenAI o3” vs “GPT-4o”/“GPT-5”. Pick one style.


Quick fix: use **grouped bars** per model (two filled bars: without/with thinking), consistent colors from the legend, add axis ticks, and label each segment explicitly. Consider a small **delta label** showing the lift from thinking.

## SWE-bench Verified

Software engineering

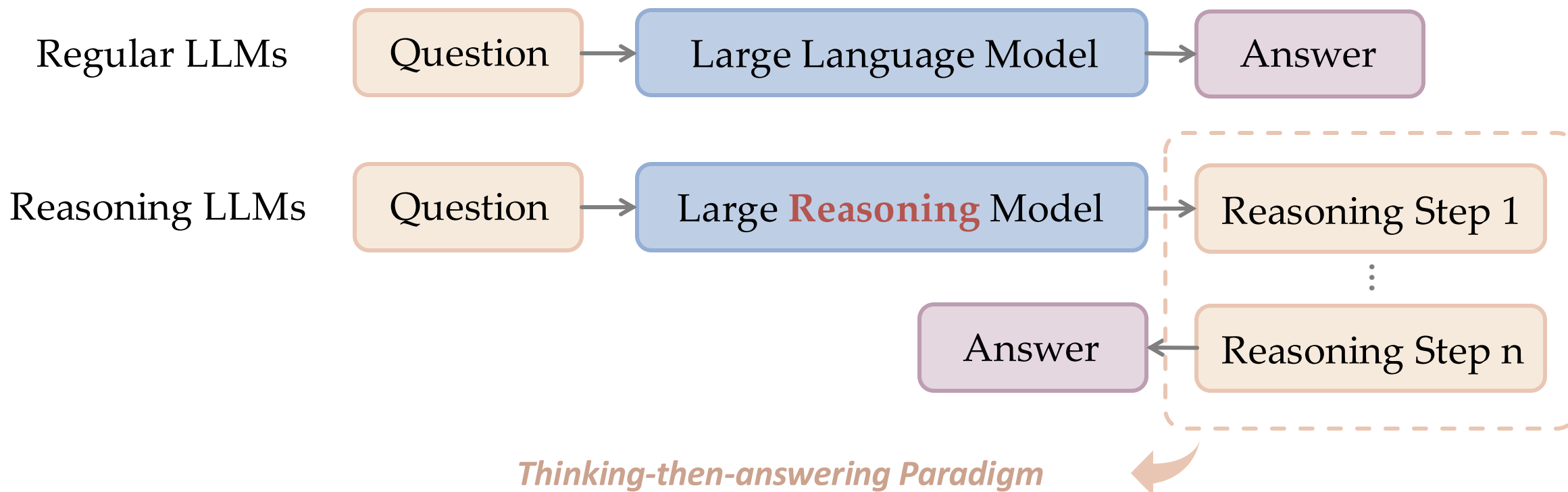
○ Without thinking   ● With thinking



GPT-5 Thinking seems oblivious to the biggest issue 

# Background: Large Reasoning Models

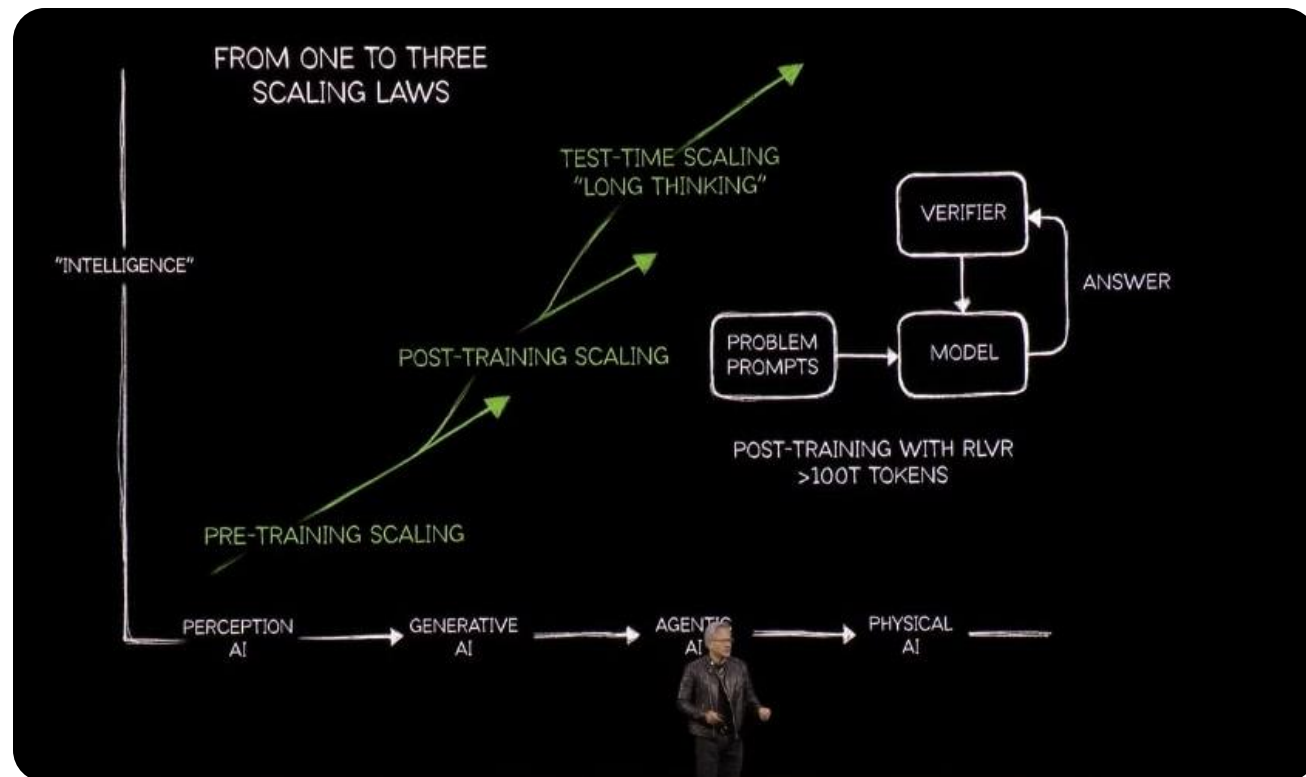
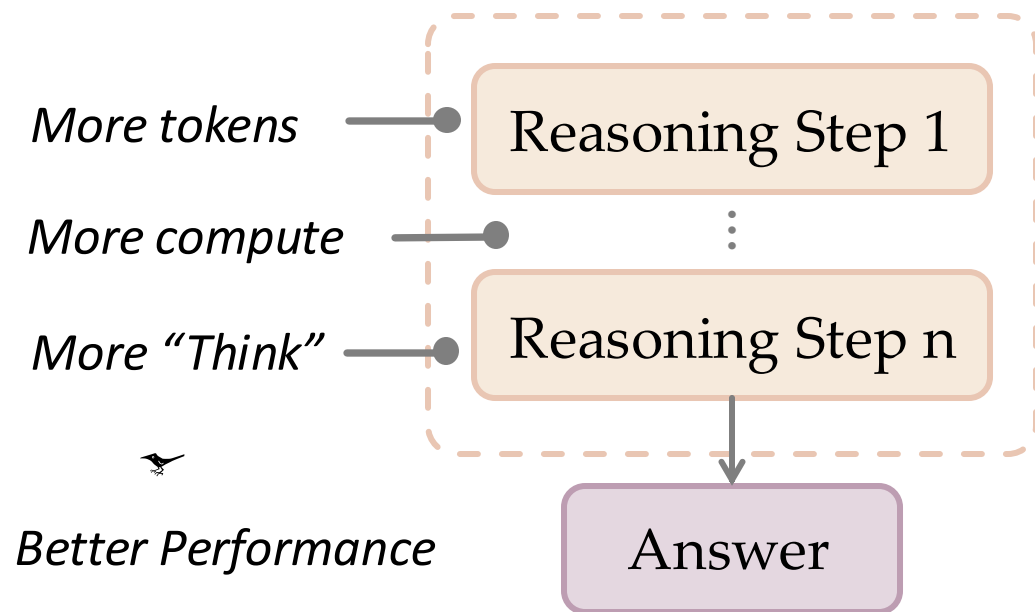
- Compared to regular LLMs, large reasoning models tend to break down a problem into smaller reasoning steps before answering a given question.



# Background: From Scaling Training to Inference

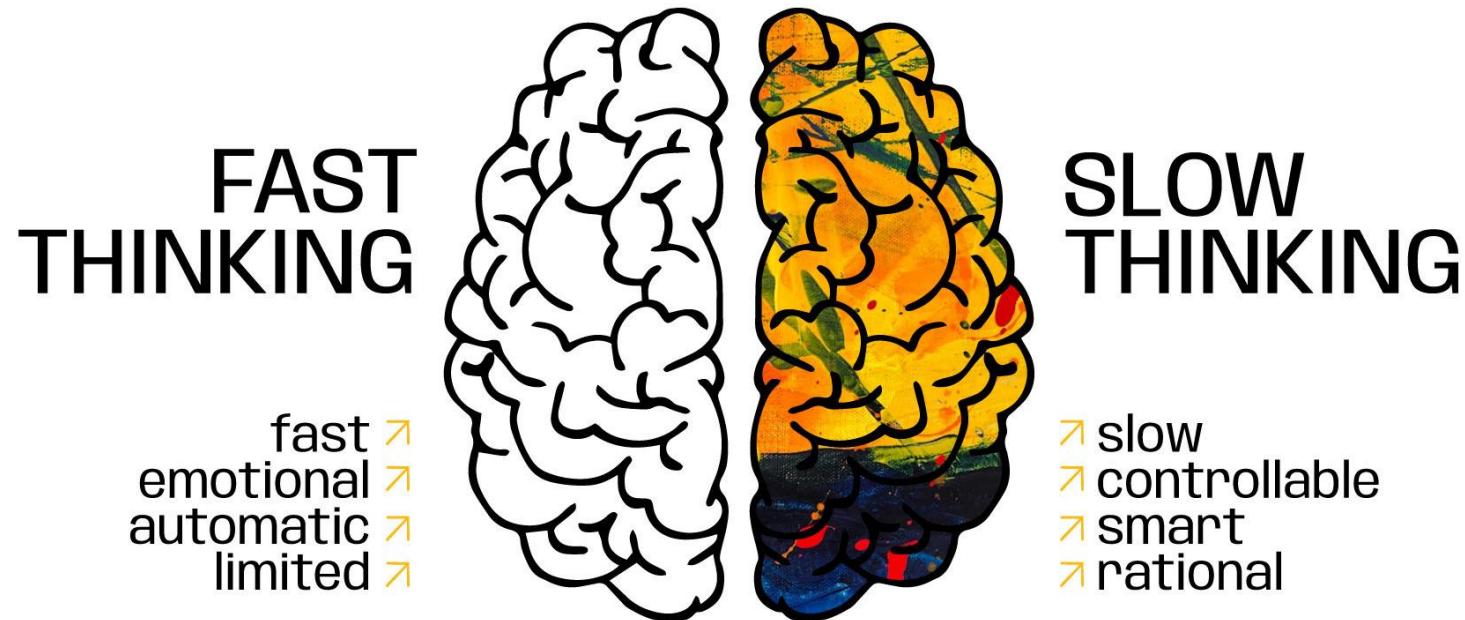
- Test-time Compute: Instead of continuously increasing pre-training budgets, test-time compute allows models to “think longer” during inference.

*beyond data* ↩



# Background: Human vs. Model Reasoning

- Humans typically think fast first and activate slow thinking through a conscious control of system-1-to-2 reasoning transitioning.
- Large reasoning models have demonstrated unprecedented progress in approaching human-like system-2 reasoning capabilities.



*Thinking, Fast and Slow.* Daniel Kahneman

# Background: Human vs. Model Reasoning

- **Challenge:** The inability of LRMs to find the optimal human-like system-1-to-2 reasoning transitioning and limited reasoning capabilities.
- Failure Example of DeepSeek-R1-Distill-Qwen-1.5B on OlympiadBench

## Example



Fail to find the optimal schedule for switching fast-slow reasoning

A chemist has three bottles, each containing a mixture of acid and water:

- bottle A contains 40g of which 10 is acid, ... She uses some of the mixture from each of the bottles to create a mixture with mass 60g of which 25 is acid. Then she mixes the remaining contents of the bottles to create a new mixture. What percentage of the new mixture is acid?

## Response from Base Model

Okay, so I've got this chemistry problem here, and I need to figure out the percentage of acid in the new mixture. Let me try to understand the problem step by step.

First, there are three bottles, each with a mixture of acid and water. The chemist is going to use some mixture from each bottle to create a new mixture that's 60 grams with 25% acid. Then, after using some from each bottle, she mixes the remaining contents to create a new mixture... </think>

The percentage of the new mixture is 18.75

# Background: Human vs. Model Reasoning

- **Challenge:** The inability of LRMs to find the optimal human-like system-1-to-2 reasoning transitioning and limited reasoning capabilities.

## Question 1


➔ *Without modulation*  
Is such automatic switching of reasoning progress determined by LRMs sufficiently reliable?



➔ *With modulation*  
Scale LRMs at test time!

# Background: Scaling LRMs at Test Time

- **Parallel Scaling:** Best of N searching strategy

- **Sequential Scaling:**  *Uniformly think more*

Monotonously **increasing** method

**s1**

How many r in raspberry?

**Question**

Let's break down the process of counting the letter 'r' in the word "raspberry" ...

\* First letter: 'r' - This is an 'r', count = 1.

\* Second letter: 'a' - Not an 'r', count remains 1 ...

\* Sixth letter: 'e' - Not an 'r', count remains 1.

\* Seventh letter: 'r' - This is an 'r', count = 2.

\* Eighth letter: 'y' - Not an 'r', count remains 2 ...

The number of 'r's in "raspberry" is 2.

**Wait**, let's re-read the question carefully. It asks "How many r in raspberry?" ... \* r - a - s - p - b - e - r - r - y ... \* First 'r' ... \* Second 'r' ... \* Third 'r' ... Count = 3 ...

**Reasoning trace**

My initial answer of 2 was incorrect due to a quick reading of the word. **Final Answer:** The final answer is **3**

**Response**

 *Uniformly think less*

Monotonously **decreasing** method

**Chain of Draft**

Standard

Answer the question directly. Do not return any preamble, explanation, or reasoning.

Chain-of-Thought

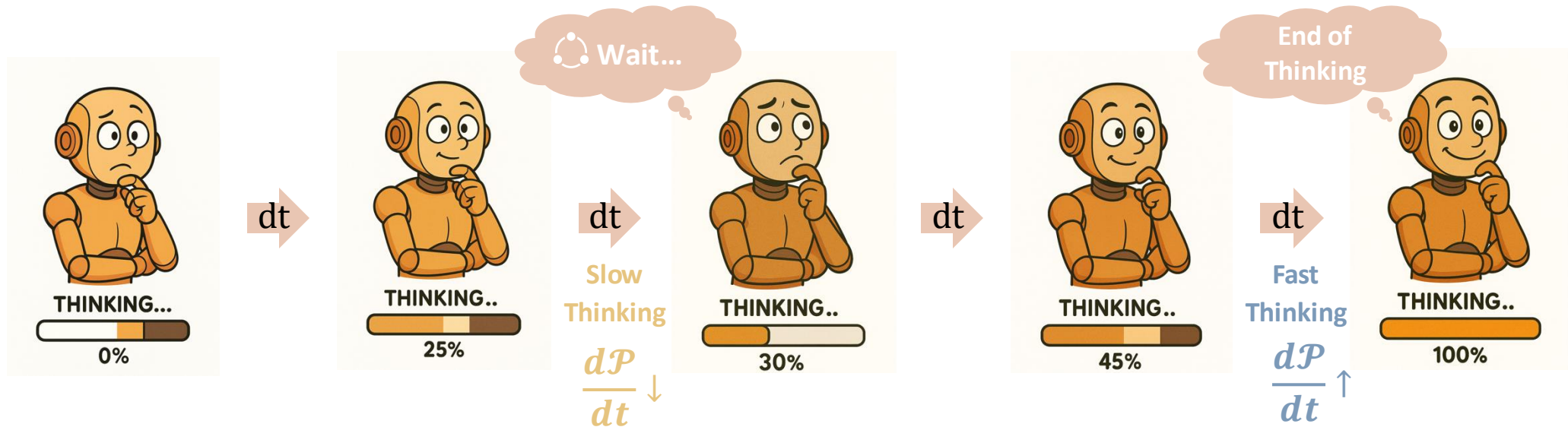
Think step by step to answer the following question. Return the answer at the end of the response after a separator ####.

Chain-of-Draft

Think step by step, but only keep a minimum draft for each thinking step, with 5 words at most. Return the answer at the end of the response after a separator ####.

# ALPHAONE: Rough Physics of Slow & Fast Thinking

**Assumption** The reasoning velocity of **slow thinking** is smaller than that of **fast thinking**.



$\mathcal{P} \in [0, 1]$ : Reasoning Progress

$dt$ : Generation Period

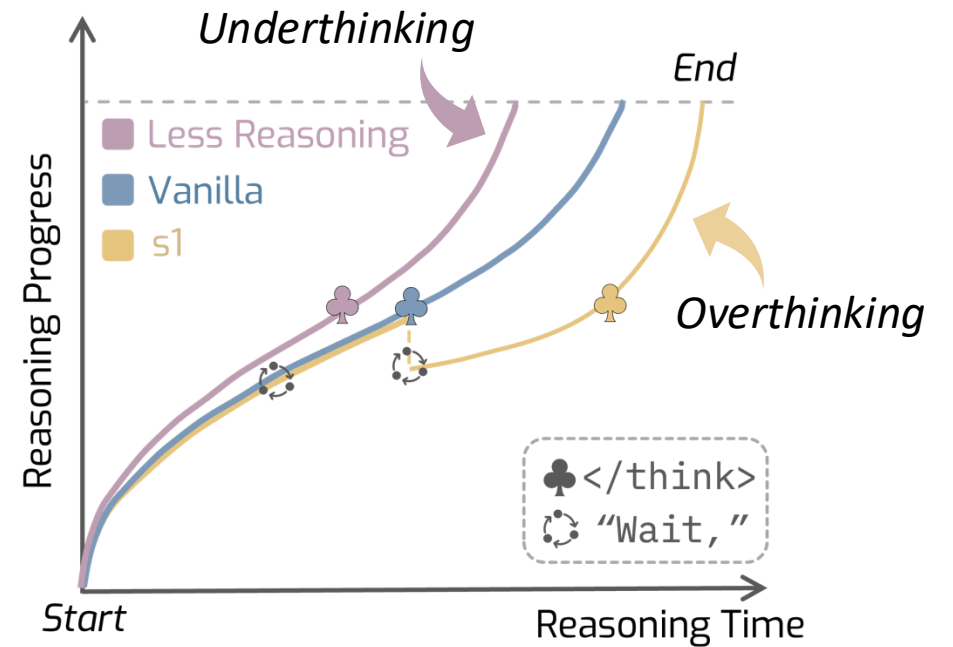
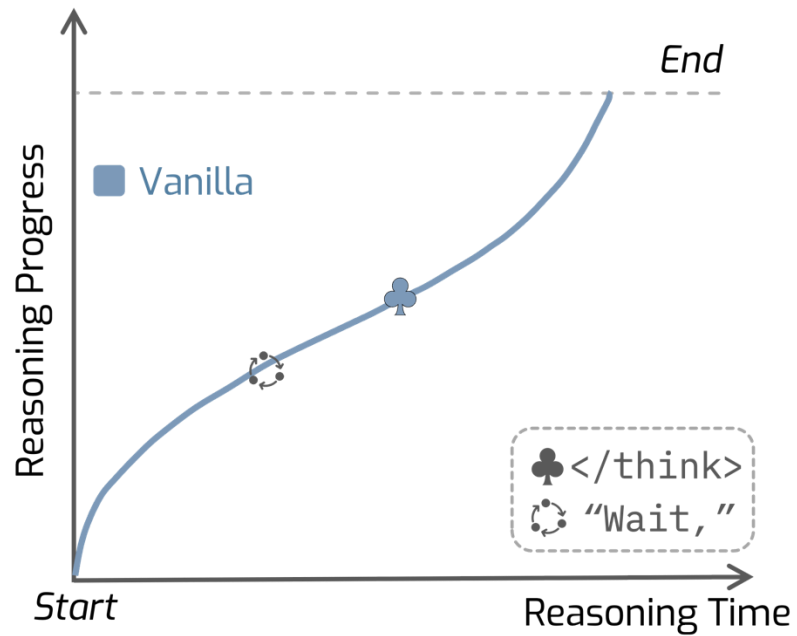
$\frac{d\mathcal{P}}{dt}$ : Reasoning Velocity

## Example:

Okay, let's try to tackle this problem step by step. Hmm, so there are 360 people in the school. ... That means the number of students taking at least one of the subjects is  $360 - 15 = 345$ . **Wait**, the total number in the union of calculus, physics, and chemistry is 345. Let me note down the given data first: ... Therefore, the number of students taking physics is 110.

# ALPHAONE: A Universal View of Test-Time Scaling

- **Parallel Scaling:** Best of N searching strategy
- **Sequential Scaling:** Monotonously **increasing** method & **decreasing** method



# ALPHAONE: A Universal View of Test-Time Scaling

- **Parallel Scaling:** Best of N searching strategy
- **Sequential Scaling:** Monotonously **increasing** method & **decreasing** method

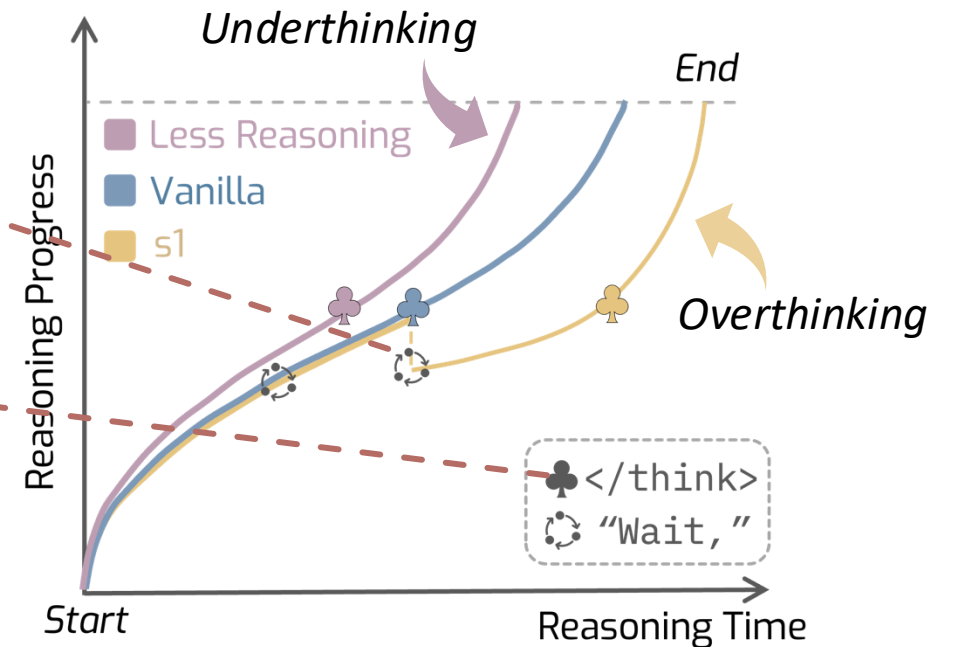
## Overthinking Example:

So the problem I need to solve is 2 plus 5, which is a very simple problem. This is what I can solve for you. **Wait**, let me think, so 2 plus 5 means doing the addition of two numbers, which are 2 and 5. So I think when you adding 2 and 5, you will get 7. \n</think>

## Underthinking Example:

Okay, so the problem given is ..... Perhaps I can solve it by ..... I can solve ..... this problem can be regarded as ..... Based on this information, I believe the answer to this problem is .....

(Wrong answer) ✘



# ALPHAONE: A Universal View of Test-Time Scaling

- **Parallel Scaling:** Best of N searching strategy
- **Sequential Scaling:** Monotonously **increasing** method & **decreasing** method

## Overthinking Example:

So the problem I need to solve is 2 plus 5, which is a very simple problem. This is what I can solve for you. **Wait**, let me think, so 2 plus 5 means doing the addition of two numbers, which are 2 and 5. So I think when you adding 2 and 5, you will get 7. \n</think>

## Underthinking Example:

Okay, so the problem given is ..... Perhaps I can solve it by ..... I can solve ..... this problem can be regarded as ..... Based on this information, I believe the answer to this problem is .....

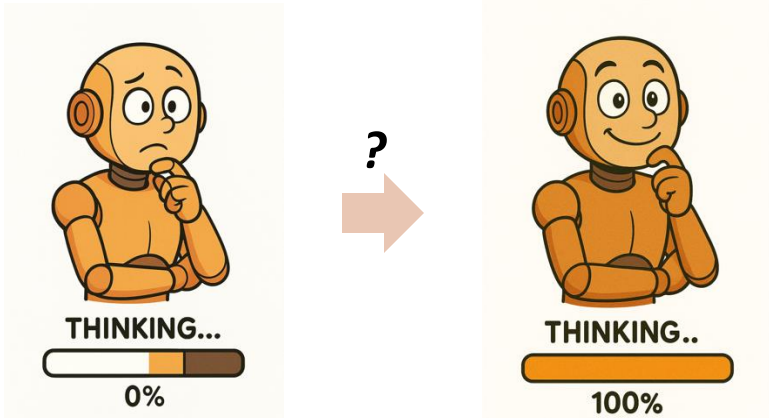
(Wrong answer) ✘

## Question 2

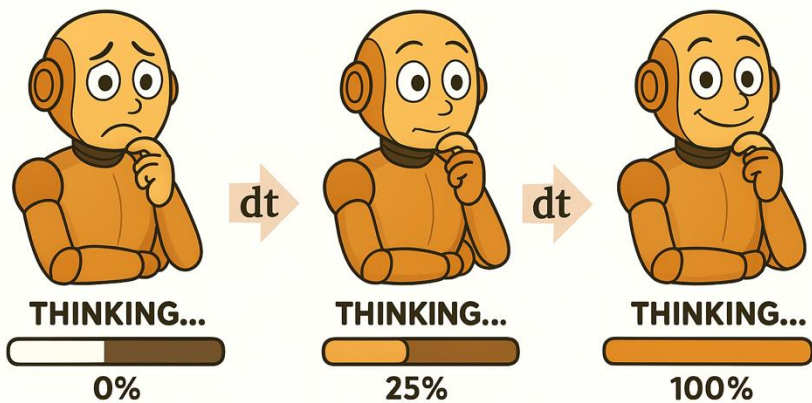
Can we modulate reasoning progress **universally**, and develop a **better thinking strategy** with it?



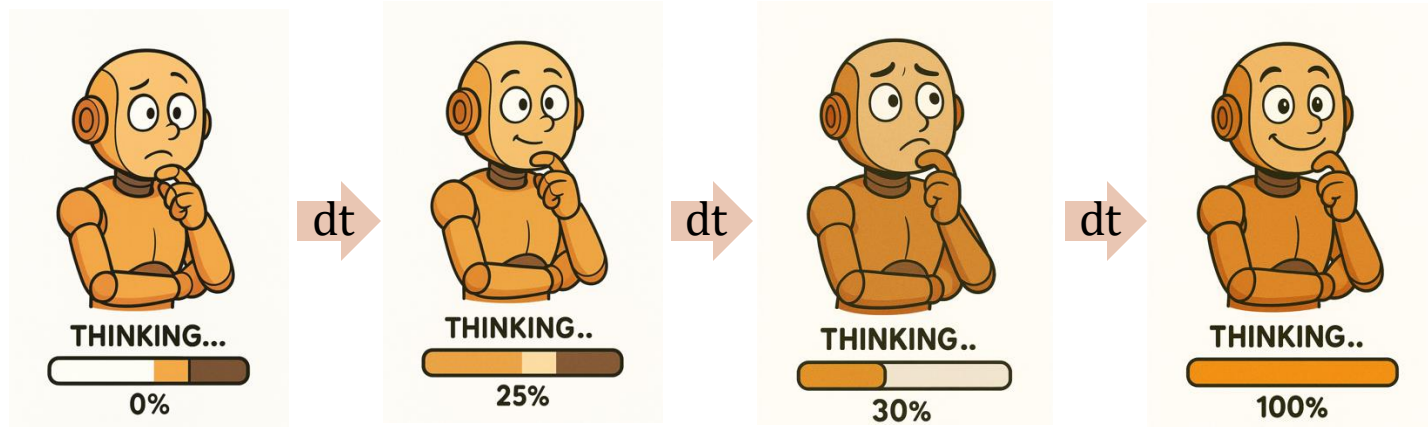
# ALPHAONE: A Universal View of Test-Time Scaling



**Thinking Phase Budgeting** *Scaling up or down thinking phase, then answer*

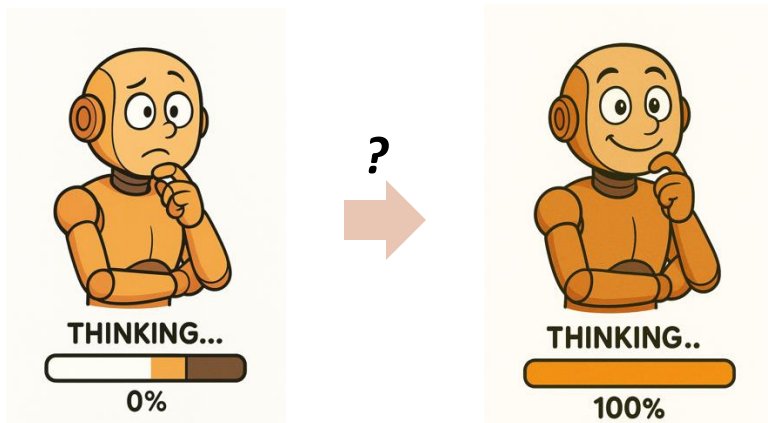


Less Thinking Phase Budget

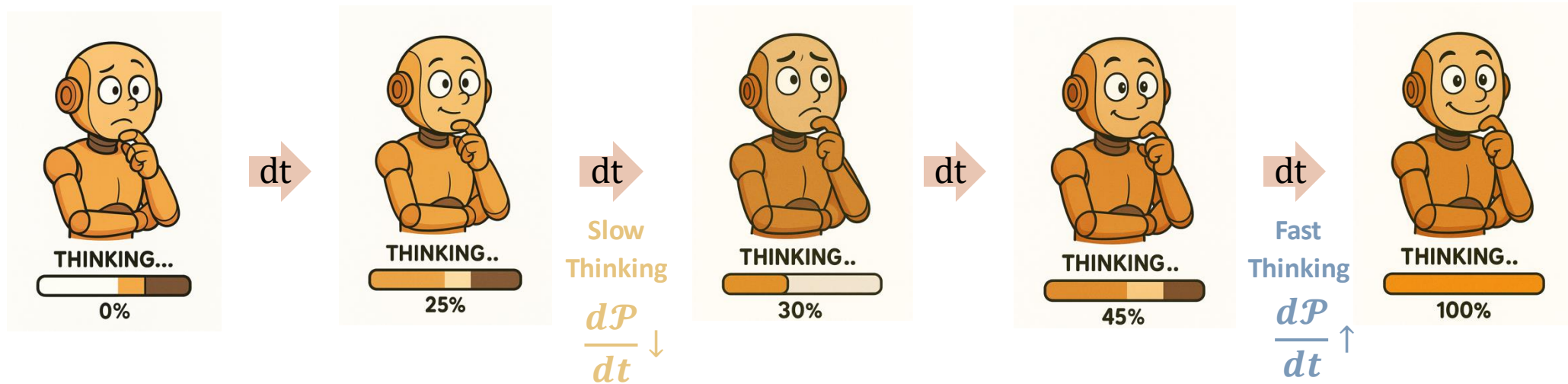


More Thinking Phase Budget

# ALPHAONE: A Universal View of Test-Time Scaling

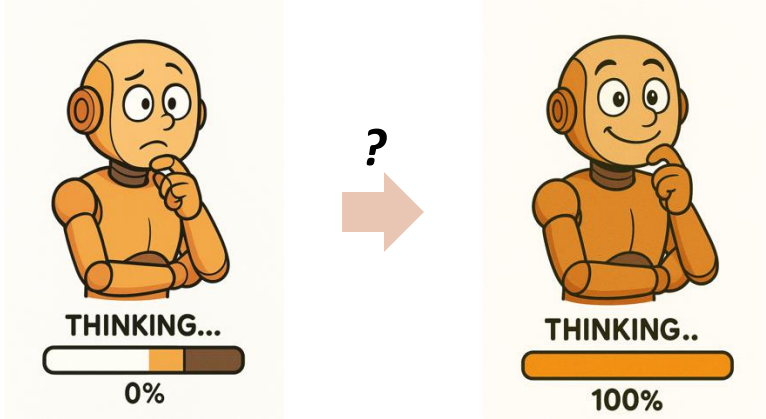


**Slow Thinking Scheduling** *Modulating slow-fast thinking transitioning arbitrarily.*

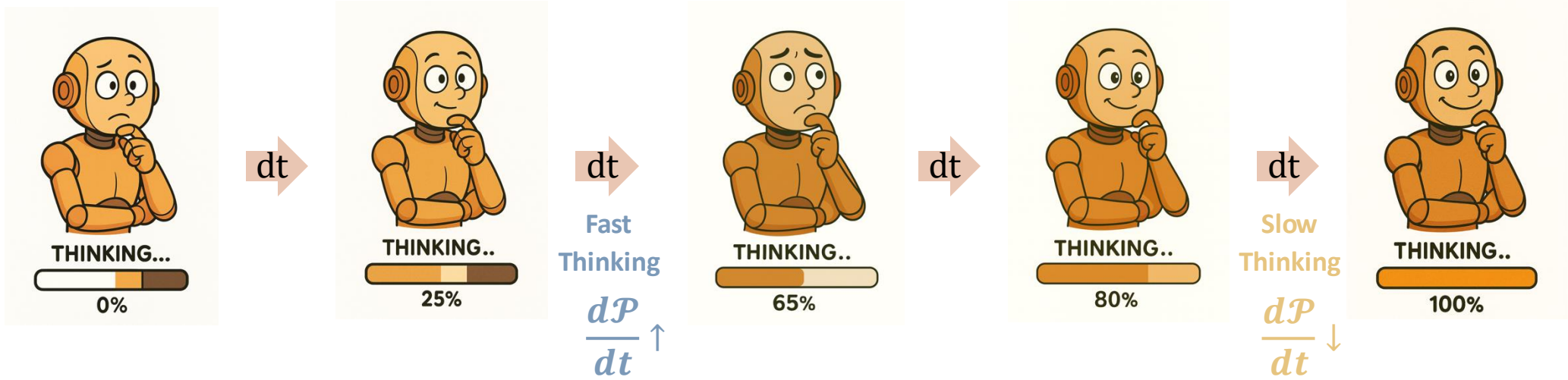


*Slow, then Fast*

# ALPHAONE: A Universal View of Test-Time Scaling



Slow Thinking Scheduling *Modulating slow-fast thinking transitioning arbitrarily.*



*Fast, then Slow*

# ALPHAONE: $\alpha$ Moment for Universal Modulation

**$\alpha$  Moment** Scales *thinking phase* by  $\alpha \times$

Given the average thinking phase token length (heuristic)  $\bar{N}_{\text{think}} > 0$  and  $\alpha > 0$ , the scaled thinking phase token length becomes  $T_m = \alpha \bar{N}_{\text{think}}$ .

- Pre  $\alpha$  Moment Modulation**

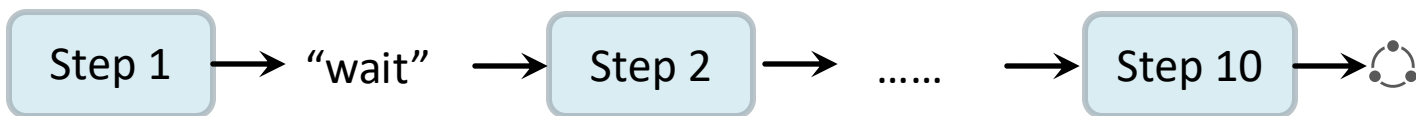
Slow Thinking Activation: “/n/n” + “wait”

Stochastic Reasoning Transitioning:

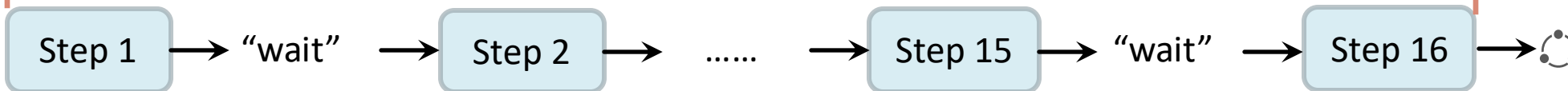
Bernoulli( $\mathbf{p}_{\text{wait}}$ )

$$\mathbf{p}_{\text{wait}} := \mathcal{S}(t), t = 0, 1, \dots, T_m.$$

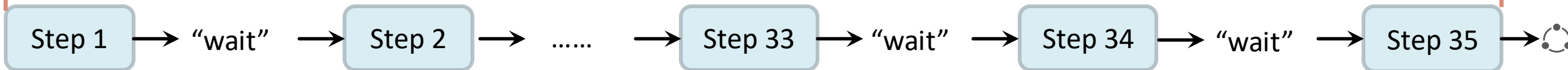
Thinking Phase Budget  $\alpha = 1$



Thinking Phase Budget  $\alpha = 1.4$



Thinking Phase Budget  $\alpha = 1.8$



# ALPHAONE: $\alpha$ Moment for Universal Modulation

**$\alpha$  Moment** Scales *thinking phase* by  $\alpha \times$

Given the average thinking phase token length (heuristic)  $\bar{N}_{\text{think}} > 0$  and  $\alpha > 0$ , the scaled thinking phase token length becomes  $T_m = \alpha \bar{N}_{\text{think}}$ .

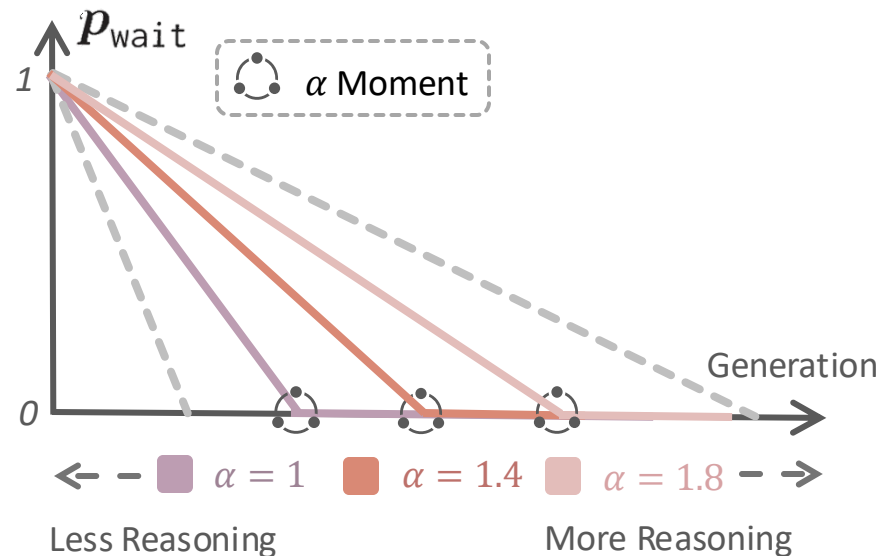
- **Pre  $\alpha$  Moment Modulation**

Slow Thinking Activation: “/n/n” + “wait”

Stochastic Reasoning Transitioning:

Bernoulli( $\mathbf{p}_{\text{wait}}$ )

$$\mathbf{p}_{\text{wait}} := \mathcal{S}(t), t = 0, 1, \dots, T_m.$$



# ALPHAONE: $\alpha$ Moment for Universal Modulation

**$\alpha$  Moment** Scales *thinking phase* by  $\alpha \times$

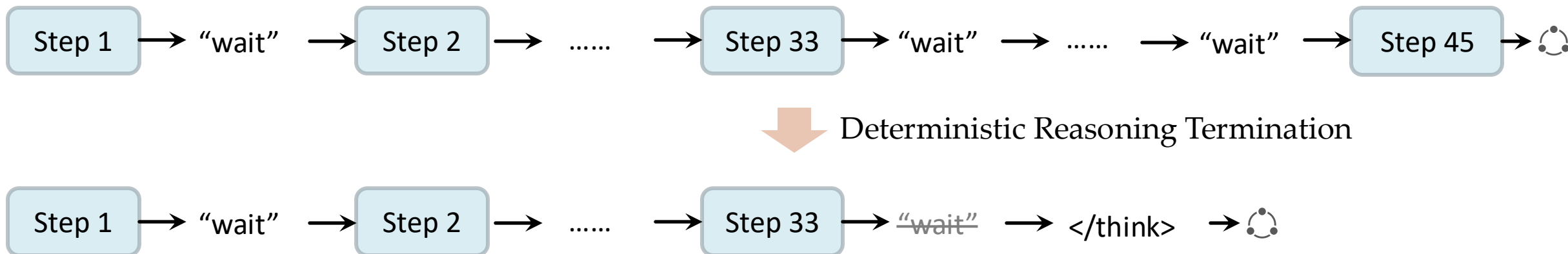
Given the average thinking phase token length (heuristic)  $\bar{N}_{\text{think}} > 0$  and  $\alpha > 0$ , the scaled thinking phase token length becomes  $T_m = \alpha \bar{N}_{\text{think}}$ .

**Slow Thinking Inertia Issue** LLMs tend to keep using slow thinking without ending

- **Post  $\alpha$  Moment Modulation**

Deterministic Reasoning Termination:

“wait”  $\Rightarrow$  “</think>”




# ALPHAONE: System Illustration

## Question

Let  $K$  be the number of sequences  $A_1, A_2, \dots, A_n$ . ... What is the remainder when  $K$  is divided by 10?

## Vanilla

Okay, so I need to figure out ... \n\n First, let me parse the problem. ... Then, ... \n\n Wait, but that can't be, because... </think> ... So, the remainder when  $K$  is divided by 10 is 0.  -----> Wait Frequency: 23

AlphaOne

Okay, so I need to figure out ... \n\n



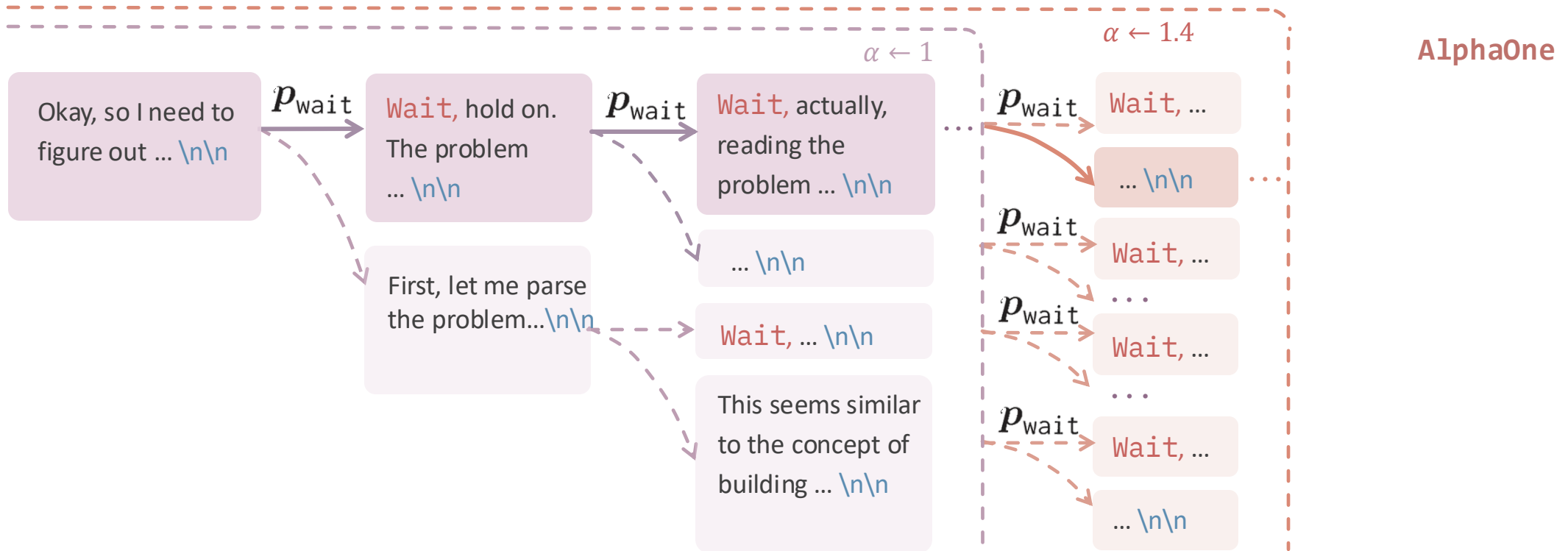
# ALPHAONE: System Illustration

## Question

Let  $K$  be the number of sequences  $A_1, A_2, \dots, A_n$ . ... What is the remainder when  $K$  is divided by 10?

## Vanilla

Okay, so I need to figure out ... \n\n First, let me parse the problem. ... Then, ... \n\n Wait, but that can't be, because... </think> ... So, the remainder when  $K$  is divided by 10 is 0. ✘ → Wait Frequency: 23





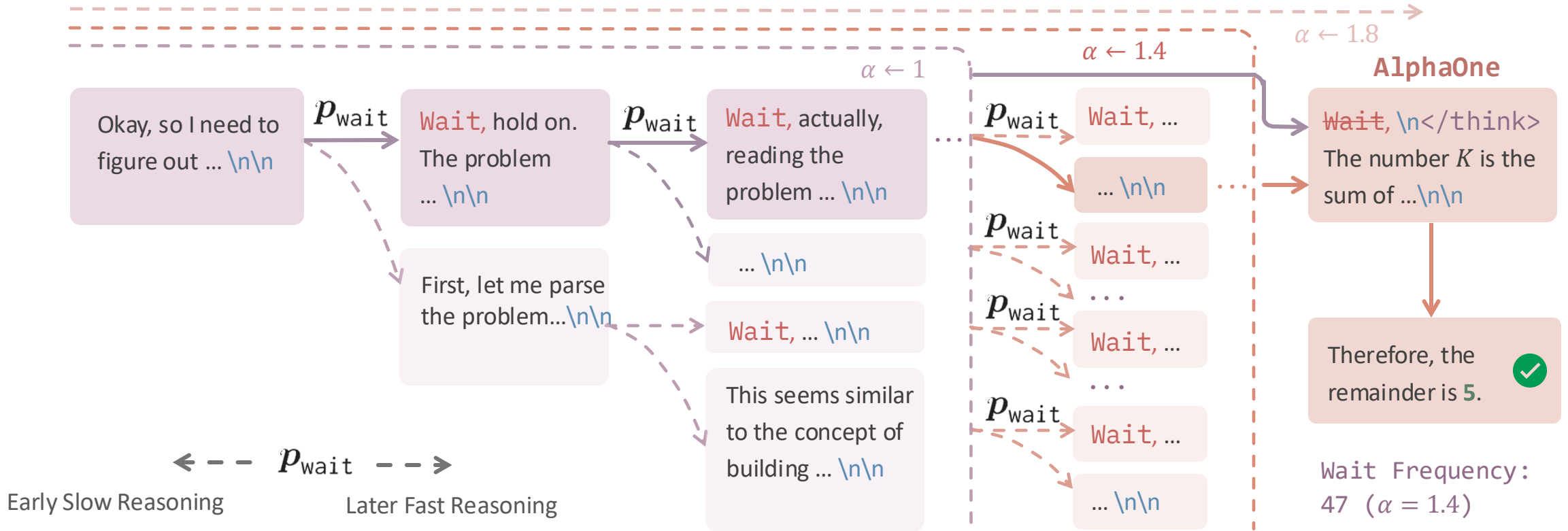
# ALPHAONE: System Illustration

## Question

Let  $K$  be the number of sequences  $A_1, A_2, \dots, A_n$ . ... What is the remainder when  $K$  is divided by 10?

## Vanilla

Okay, so I need to figure out ... \n\n First, let me parse the problem. ... Then, ... \n\n ~~Wait~~, but that can't be, because... </think> ... So, the remainder when  $K$  is divided by 10 is 0. ✘ Wait Frequency: 23



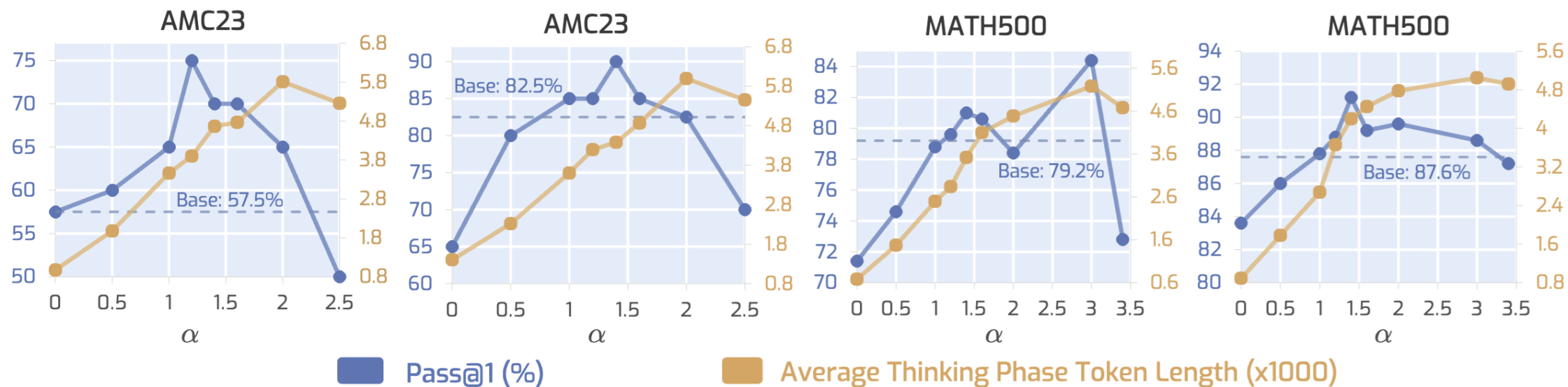
# ALPHAONE: Systematic Comparison



- How to determine  $\alpha$  that controls slow-fast thinking effectively?
- How to perform scheduling for reasoning modulation?

# ALPHAONE: Can $\alpha$ -moment scale the thinking phase budget?

- **Scaling property of  $\alpha$ :** scale  $\alpha$  from 0 to the maximum value
  - 1  $\alpha$ -moment enables a **scalable** thinking phase budgeting
  - 2 While the thinking phase is scaled up, there exists a **trade-off** between the optimal value of  $\alpha$  and the resulting reasoning accuracy



(a) DeepSeek-R1-Distill-Qwen-1.5B

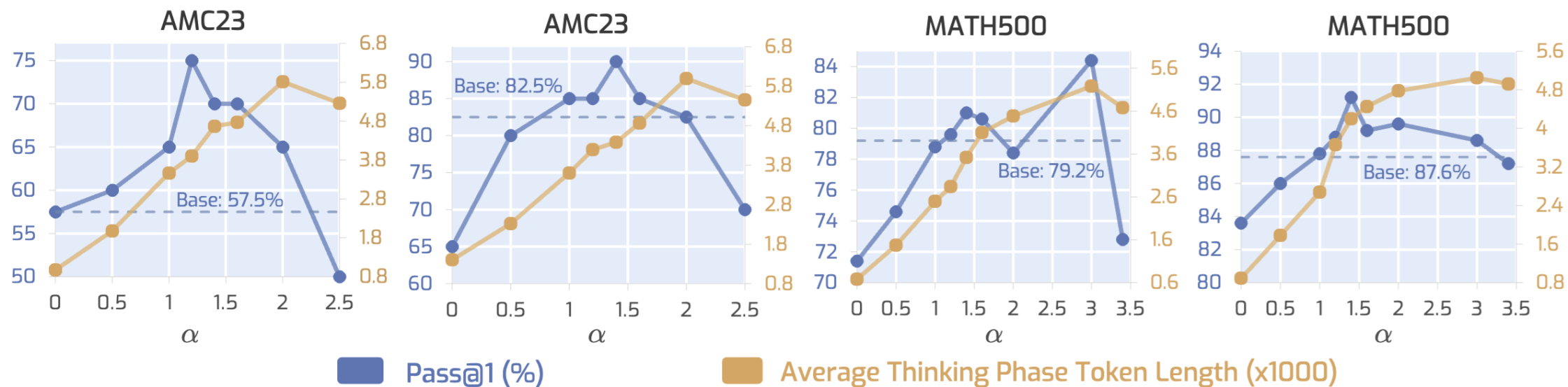
(b) DeepSeek-R1-Distill-Qwen-7B

(c) DeepSeek-R1-Distill-Qwen-1.5B

(d) DeepSeek-R1-Distill-Qwen-7B

# ALPHAONE: Can $\alpha$ -moment scale the thinking phase budget?

- Scaling property of  $\alpha$ : scale  $\alpha$  from 0 to the maximum value
- 3 The **robustness** of  $\alpha$ : Across different values of  $\alpha$ ,  $\alpha 1$  consistently outperforms the base model by a substantial margin



(a) DeepSeek-R1-Distill-Qwen-1.5B

(b) DeepSeek-R1-Distill-Qwen-7B

(c) DeepSeek-R1-Distill-Qwen-1.5B

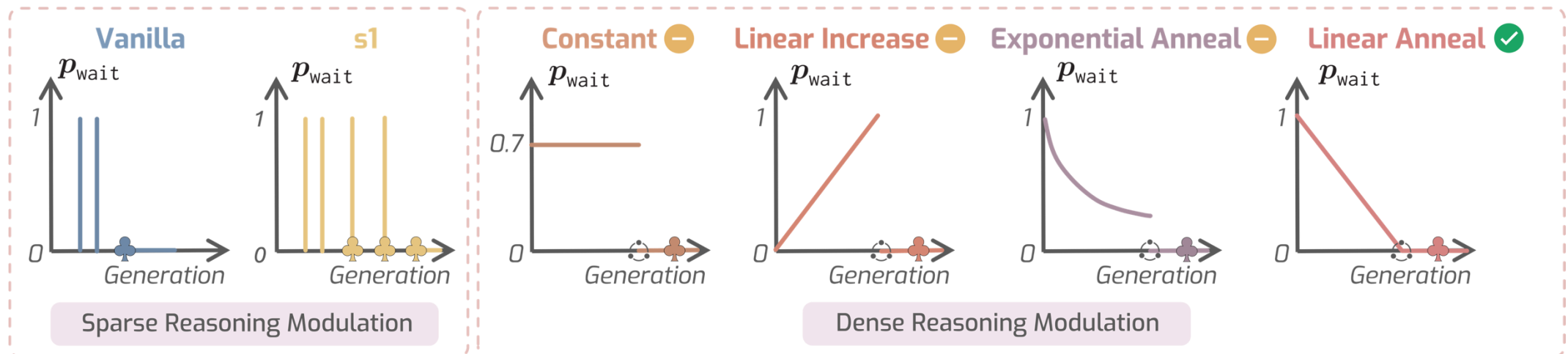
(d) DeepSeek-R1-Distill-Qwen-7B

# ALPHAONE: What scheduling strategy is better?

- Four different scheduling strategies

- 1 **Constant:**  $\mathcal{S}(t) := p_{\text{constant}}$
- 2 **Linear Increase:**  $\mathcal{S}(t) := \frac{1}{T_m}t$
- 3 **Exponential Anneal:**  $\mathcal{S}(t) := \exp(-\gamma t)$
- 4 **Linear Anneal:**  $\mathcal{S}(t) := -\frac{1}{T_m}t + 1$

$T_m = \alpha \bar{N}_{\text{think}}$  represent the timestamp of  $\alpha$ -moment,  $t = \{0, 1, \dots, T_m\}$ , and  $\gamma > 0$

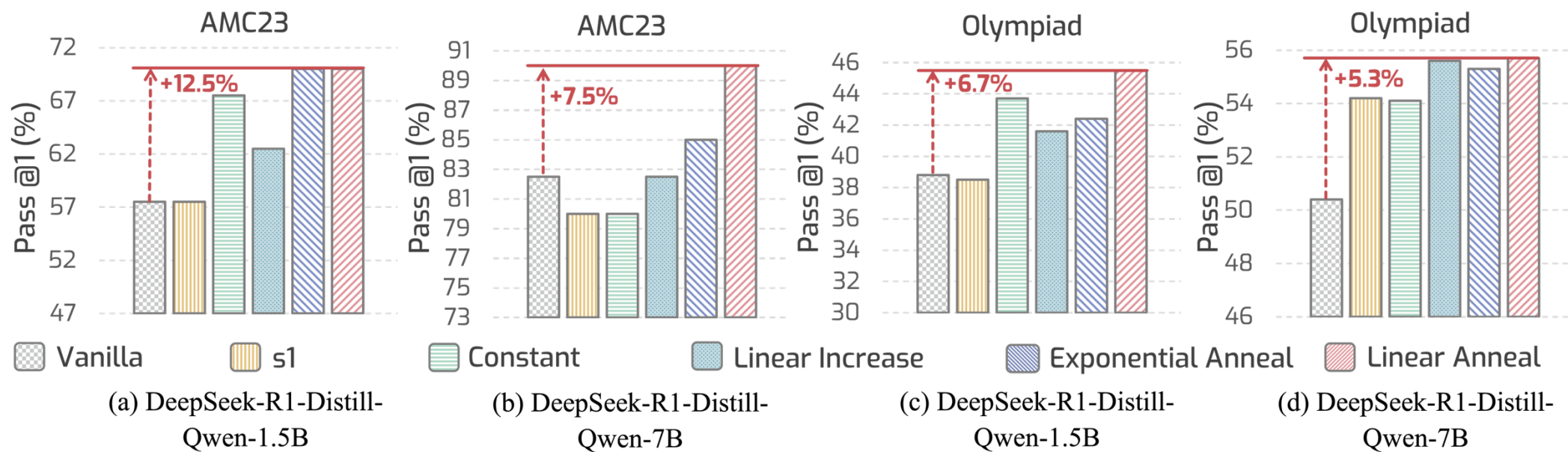


# ALPHAONE: What scheduling strategy is better?

- Linear annealing consistently yields the highest reasoning accuracy.

## Finding 1

Slow thinking first, then fast thinking, leads to better LRM reasoning



# ALPHAONE: Systematic Comparison

- $\alpha 1$  consistently yields a **higher** problem-solving accuracy.
- While  $\alpha 1$  modulates reasoning densely without restrictions on reducing the thinking budget, it achieves more **efficient reasoning** than baselines.

Method	MATHEMATICAL								CODING		SCIENCE		$\bar{\Delta}_{P@1}$
	AIME24		AMC23		Minerva		MATH500		LiveCode		Olympiad		
	P@1	#Tk	P@1	#Tk	P@1	#Tk	P@1	#Tk	P@1	#Tk	P@1	#Tk	
<i>DeepSeek-R1-Distill-Qwen-1.5B</i>													
BASE	23.3	7280	57.5	5339	32.0	4935	79.2	3773	17.8	6990	38.8	5999	N/A
s1*	26.7 <sub>+3.4</sub>	7798	57.5 <sub>+0.0</sub>	6418	31.6 <sub>-0.4</sub>	5826	78.2 <sub>-1.0</sub>	4733	17.0 <sub>-0.8</sub>	7025	38.5 <sub>-0.3</sub>	6673	+0.15
CoD	<b>30.0<sub>+6.7</sub></b>	6994	65.0 <sub>+7.5</sub>	5415	29.0 <sub>-3.0</sub>	<b>4005</b>	<b>81.4<sub>+2.2</sub></b>	<b>3136</b>	20.3 <sub>+2.5</sub>	6657	40.6 <sub>+1.8</sub>	5651	+2.95
$\alpha 1$ (Ours)	<b>30.0<sub>+6.7</sub></b>	<b>5916</b>	<b>70.0<sub>+12.5</sub></b>	<b>4952</b>	<b>34.2<sub>+2.2</sub></b>	4586	81.0 <sub>+1.8</sub>	3852	<b>24.8<sub>+7.0</sub></b>	<b>5426</b>	<b>45.5<sub>+6.7</sub></b>	<b>4944</b>	<b>+6.15</b>
<i>DeepSeek-R1-Distill-Qwen-7B</i>													
BASE	46.7	6648	82.5	4624	40.4	4191	87.6	3239	43.5	5885	50.4	5385	N/A
s1*	46.7 <sub>+0.0</sub>	7295	80.0 <sub>-2.5</sub>	5673	42.3 <sub>+1.9</sub>	6510	<b>92.8<sub>+5.2</sub></b>	5848	44.0 <sub>+0.5</sub>	5979	54.2 <sub>+3.8</sub>	6007	+1.48
CoD	43.3 <sub>-3.4</sub>	<b>6078</b>	87.5 <sub>+5.0</sub>	<b>3594</b>	<b>43.4<sub>+3.0</sub></b>	<b>2142</b>	88.8 <sub>+1.2</sub>	<b>2094</b>	45.0 <sub>+1.5</sub>	5593	53.5 <sub>+3.1</sub>	<b>4520</b>	+1.73
$\alpha 1$ (Ours)	<b>50.0<sub>+3.3</sub></b>	6827	<b>90.0<sub>+7.5</sub></b>	4397	42.3 <sub>+1.9</sub>	4124	91.2 <sub>+3.6</sub>	4337	<b>49.8<sub>+6.3</sub></b>	5067	<b>55.7<sub>+5.3</sub></b>	4883	<b>+4.65</b>
<i>Qwen QwQ-32B</i>													
BASE	40.0	4058	77.5	2901	<b>47.8</b>	2199	90.2	1951	67.0	5092	53.6	3230	N/A
s1*	43.3 <sub>+3.3</sub>	4221	77.5 <sub>+0.0</sub>	3068	46.7 <sub>-1.1</sub>	2433	<b>90.8<sub>+0.6</sub></b>	2218	66.5 <sub>-0.5</sub>	5260	55.1 <sub>+1.5</sub>	3454	+0.63
CoD	46.7 <sub>+6.7</sub>	3959	80.0 <sub>+2.5</sub>	2400	47.4 <sub>-0.4</sub>	1464	90.6 <sub>+0.4</sub>	<b>1421</b>	66.8 <sub>-0.2</sub>	<b>4984</b>	<b>57.2<sub>+3.6</sub></b>	2844	+2.10
$\alpha 1$ (Ours)	<b>53.3<sub>+13.3</sub></b>	<b>3141</b>	<b>87.5<sub>+10.0</sub></b>	<b>2286</b>	46.0 <sub>-1.8</sub>	<b>1441</b>	89.4 <sub>-0.8</sub>	1668	<b>75.8<sub>+8.8</sub></b>	5824	56.1 <sub>+2.5</sub>	<b>2504</b>	<b>+5.33</b>

## ALPHAONE: Does $\alpha 1$ scale more efficiently?

- **REP (Reasoning Efficiency-Performance) Metric:** quantitatively evaluate how different methods trade off reasoning efficiency and accuracy
- Higher REP indicates stronger performance with better reasoning efficiency.

*Reasoning accuracy of the evaluated method*

$$\mathcal{F}_{\text{REP}}(\mathcal{A}_{\text{method}}; \mathcal{A}_{\text{base}}, T_{\text{norm}}) = \frac{\mathcal{A}_{\text{method}} - \mathcal{A}_{\text{base}}}{T_{\text{norm}}}$$

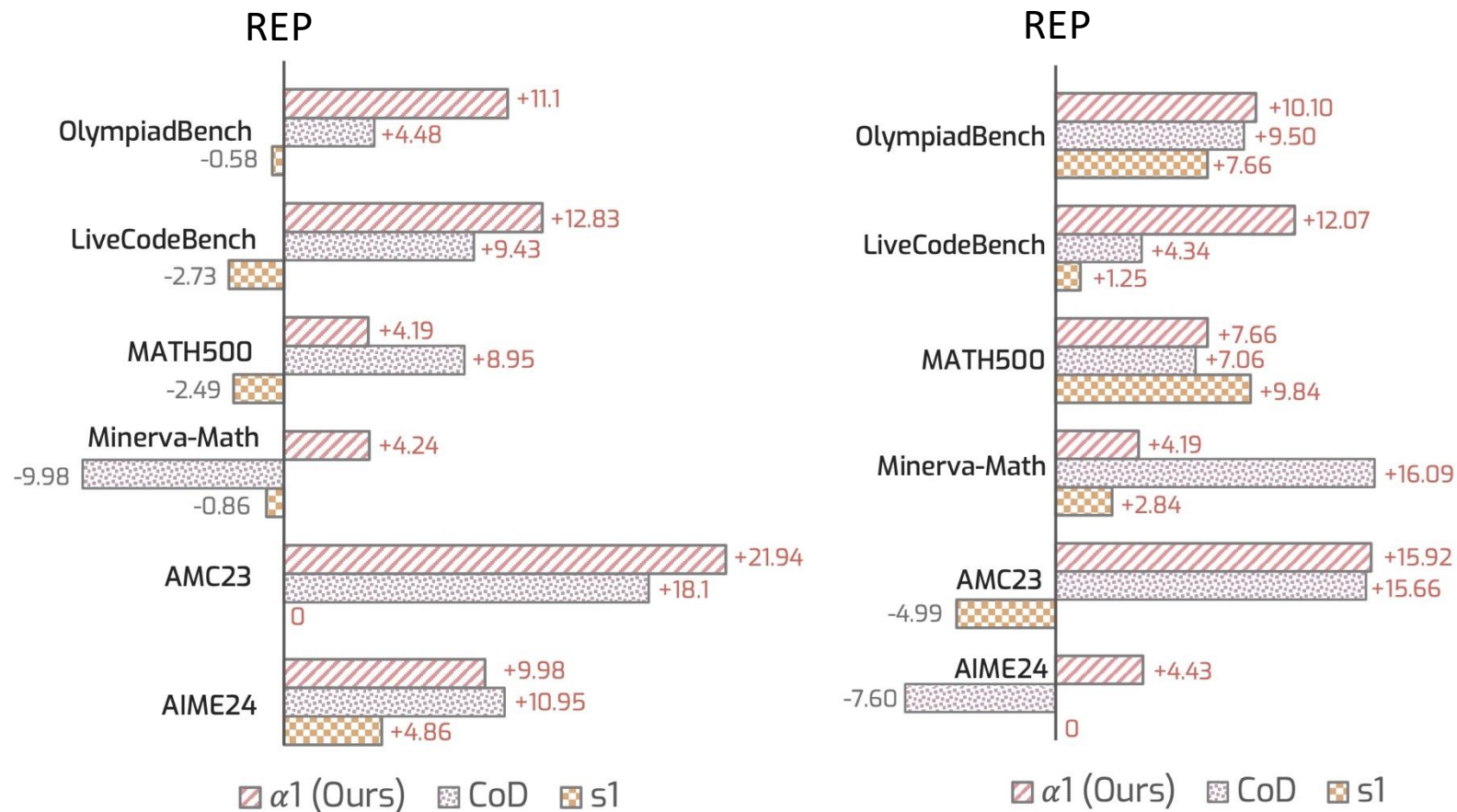
*Normalized thinking phase token length*

- $\alpha 1$  achieves a favorable balance between reasoning performance and efficiency.

# ALPHAONE: Does $\alpha 1$ scale more efficiently?

## Finding 2

Slow thinking can bring efficient test-time scaling.

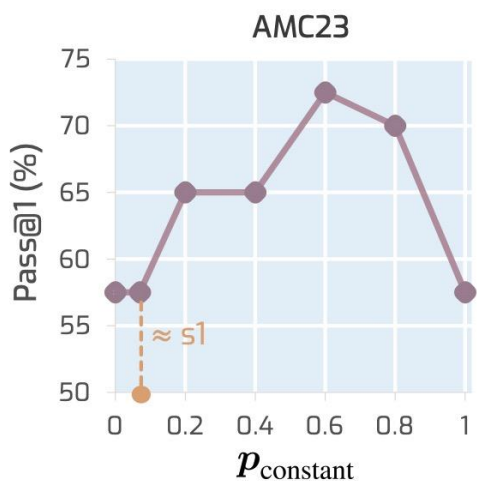


# ALPHAONE: How frequent should slow thinking transitioning be?

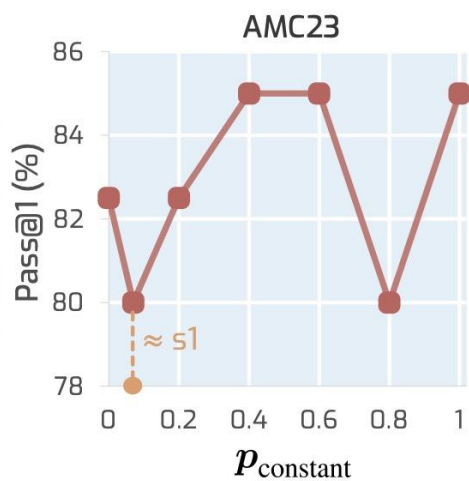
- Scaling property of “wait” frequency under constant scheduling

## Finding 3

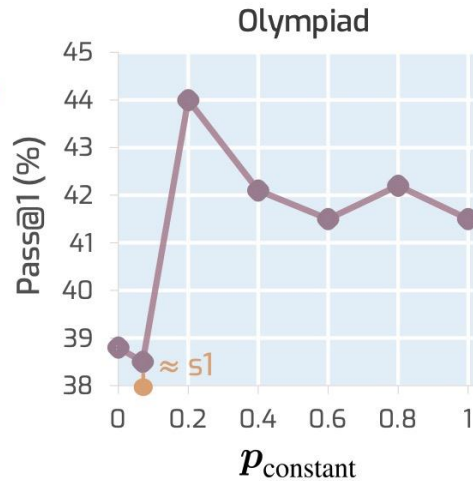
Slow thinking scheduling in high frequency is helpful.



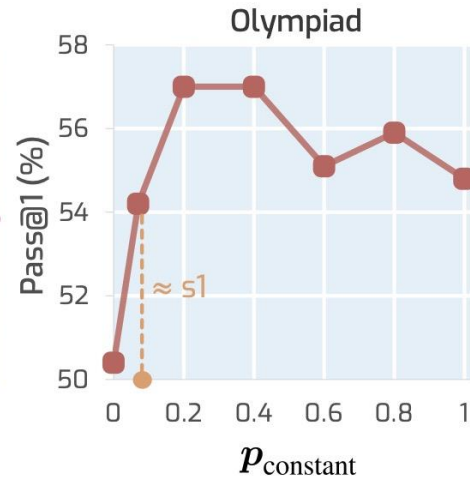
(a) DeepSeek-R1-Distill-Qwen-1.5B



(b) DeepSeek-R1-Distill-Qwen-7B



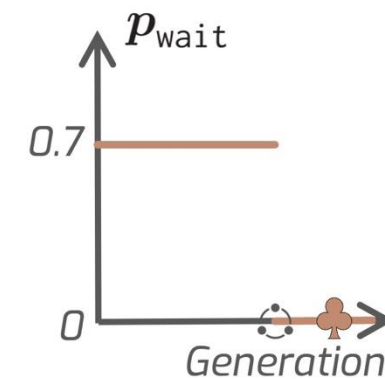
(c) DeepSeek-R1-Distill-Qwen-1.5B



(d) DeepSeek-R1-Distill-Qwen-7B

$$\mathcal{S}(t) := p_{\text{constant}}$$

Constant —



constant scheduling

# ALPHAONE: Is post- $\alpha$ moment modulation necessary?

- Pre- $\alpha$  moment modulation of slow thinking is insufficient

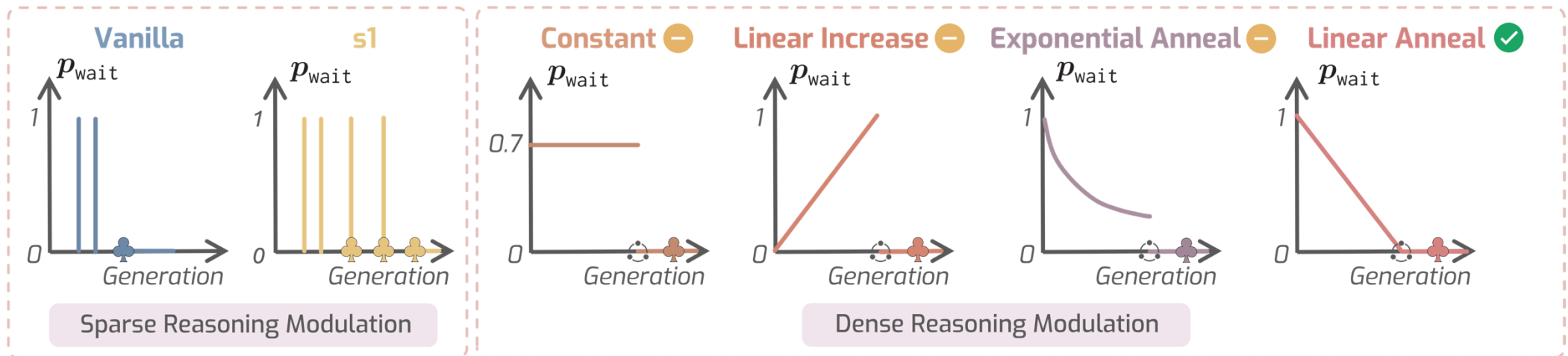
*slow thinking inertia* 

- $\alpha 1$  successfully ends in a fast thinking with post- $\alpha$  moment modulation

Method	Post- $\alpha$ Moment Modulation	AIME24		AMC23	
		P@1	#Tk	P@1	#Tk
<i>DeepSeek-R1-Distill-Qwen-1.5B</i>					
BASE	N/A	23.3	7280	57.5	5339
$\alpha 1$ (Ours)	×	26.7	7929	47.5	6903
$\alpha 1$ (Ours)	✓	<b>30.0</b>	<b>5916</b>	<b>70.0</b>	<b>4951</b>
<i>DeepSeek-R1-Distill-Qwen-7B</i>					
BASE	N/A	38.8	<b>5999</b>	82.5	4624
$\alpha 1$ (Ours)	×	30.0	7666	75.0	5878
$\alpha 1$ (Ours)	✓	<b>50.0</b>	6826	<b>90.0</b>	<b>4397</b>

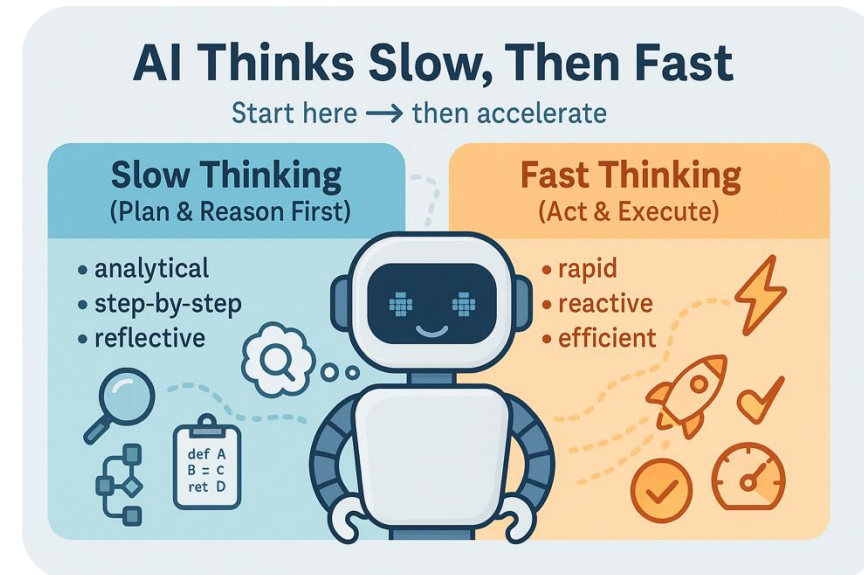
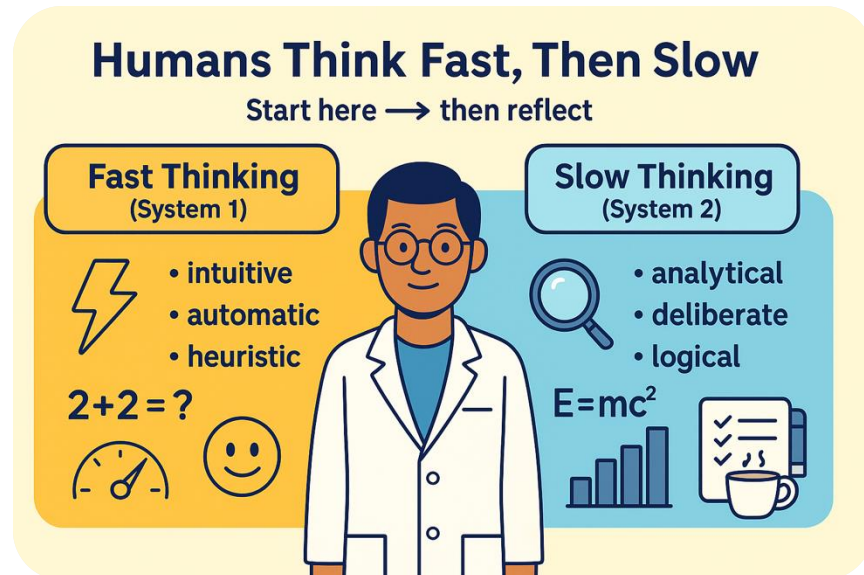
# ALPHAONE: Beyond the Method

- 1  $\alpha 1$  is a universal framework for reasoning modulation: It unifies and generalizes existing monotonic scaling methods by enabling flexible and dense slow-to-fast reasoning modulation.



# ALPHAONE: Beyond the Method

- $\alpha 1$  challenges prevailing reasoning paradigms:** The reasoning models fail to find the optimal scheduling for switching fast-slow reasoning without modulation.
- $\alpha 1$  utilizes a simple method to motivate an eternal problem:** The optimal test-time reasoning modulation is still unknown.



# The Future of ALPHAOONE



**More sophisticated slow-fast thinking scheduling:** Modulate reasoning progress during both training and inference, or learn a separate progress modulation model aligned with human preferences



**Transitioning-token-agnostic modulation:** While  $\alpha 1$  supports flexible token choices, removing the dependency on transitioning tokens altogether could further enhance generalization

# The Future of ALPHAOONE



**$\alpha$ 1-style reasoning modulation with RL:** Utilize reinforcement learning to learn an adaptive reasoning modulation in  $\alpha$ 1 fashion. For instance,  $\alpha$ -moment can be adaptively sampled from a subnetwork, and the reasoning scheduling strategy can be adaptively selected when facing different problems.



**$\alpha$ 1 as a sampling approach:** utilize  $\alpha$ 1 to modulate the reasoning progress when sampling rollouts. This brings more reasoning pattern diversity of classical methods like GRPO.

# The Future of ALPHAOONE



**Multimodal reasoning with multimodal LLMs:** Extend  $\alpha 1$  framework to multimodal domain, fostering synergistic multimodal comprehension and creation.

# ALPHAONE

- **Takeaways**
  - We present AlphaOne ( $\alpha 1$ ), a universal framework for modulating reasoning progress in large reasoning models (LRMs) at test time.



Paper



Website



GitHub





**Thank You!**

**Q & A**

**NVIDIA, Aug 15, 2025**



**UNIVERSITY OF  
ILLINOIS**  
URBANA-CHAMPAIGN

**Berkeley**

