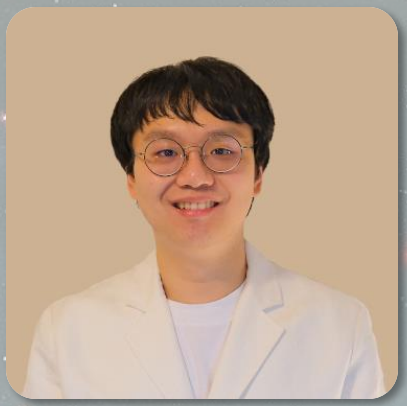
AlphaOne
Reasoning Models Thinking Slow and Fast at Test Time







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

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN Berkeley

Background: Large Reasoning Models

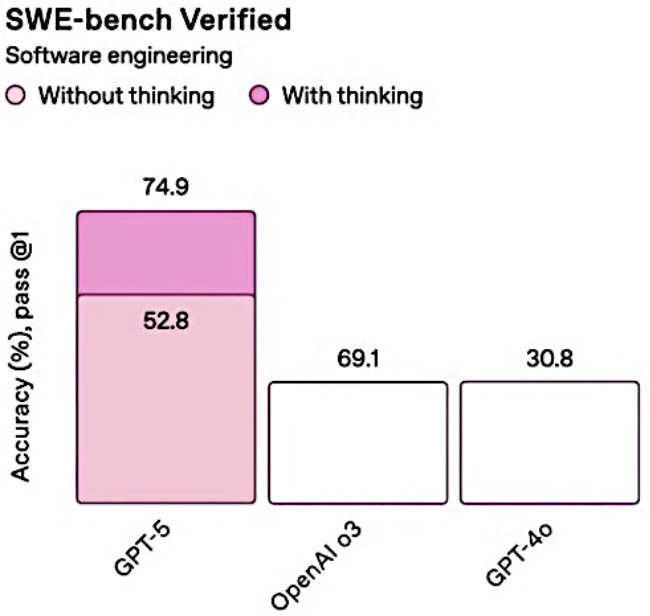
Notice anything?

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.

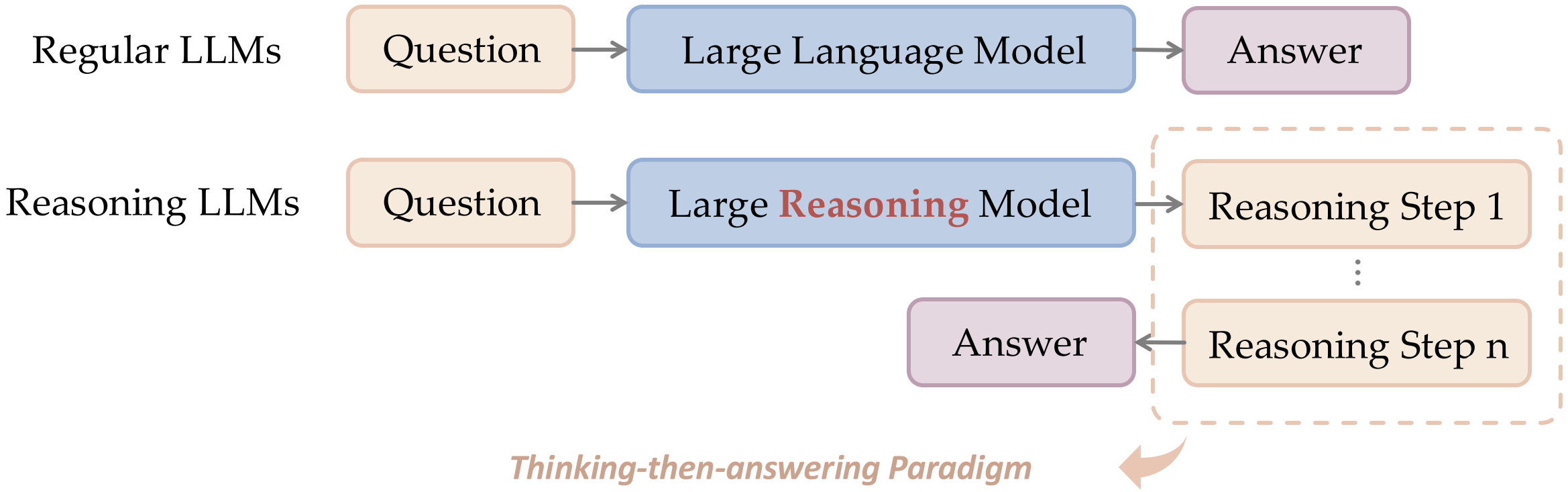
Quick fix: use **grouped bars** per model (two filled bars: without/with thinking), consistent colors from the legend, add axis ticks, and label each segment explicitly. Consider a small **delta label** showing the lift from thinking.



GPT-5 Thinking seems oblivious to the biggest issue 😈

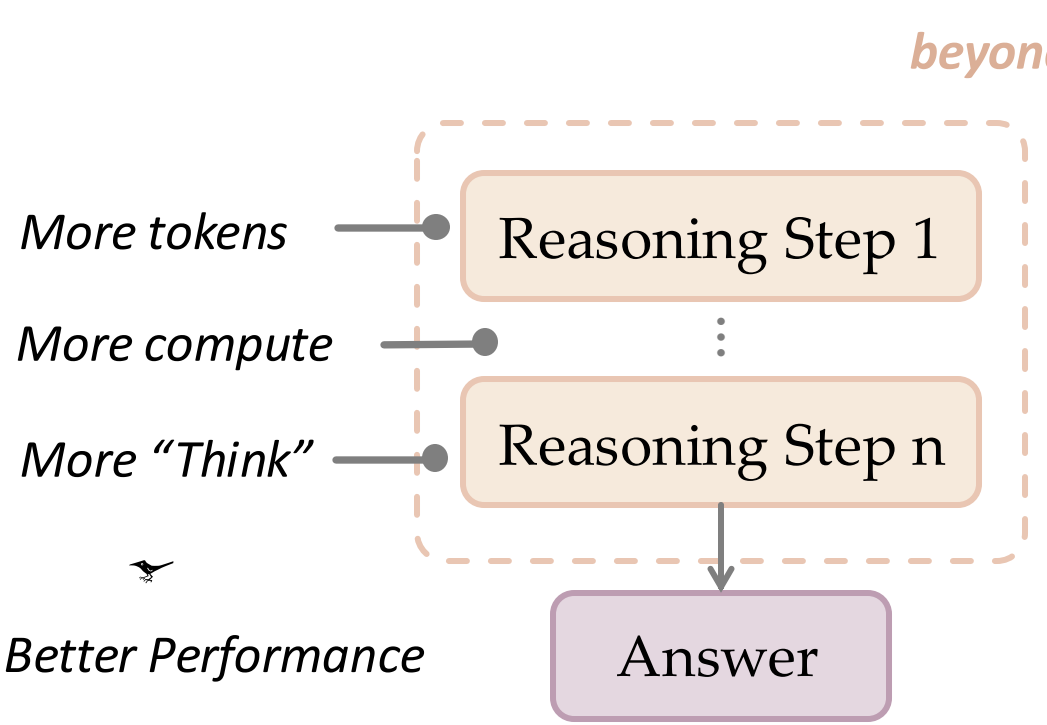
Background: Large Reasoning Models

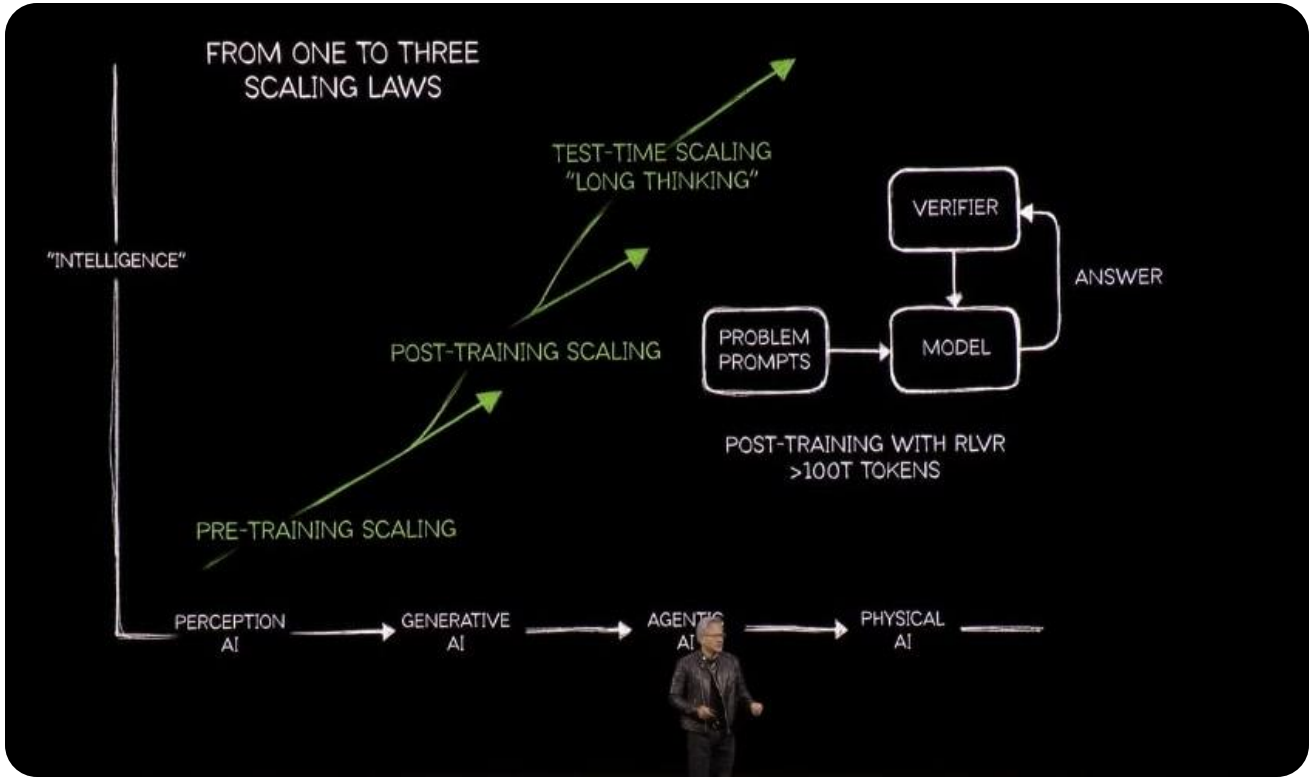
- Compared to regular LLMs, large reasoning models tend to break down a problem into smaller reasoning steps before answering a given question.



Background: From Scaling Training to Inference

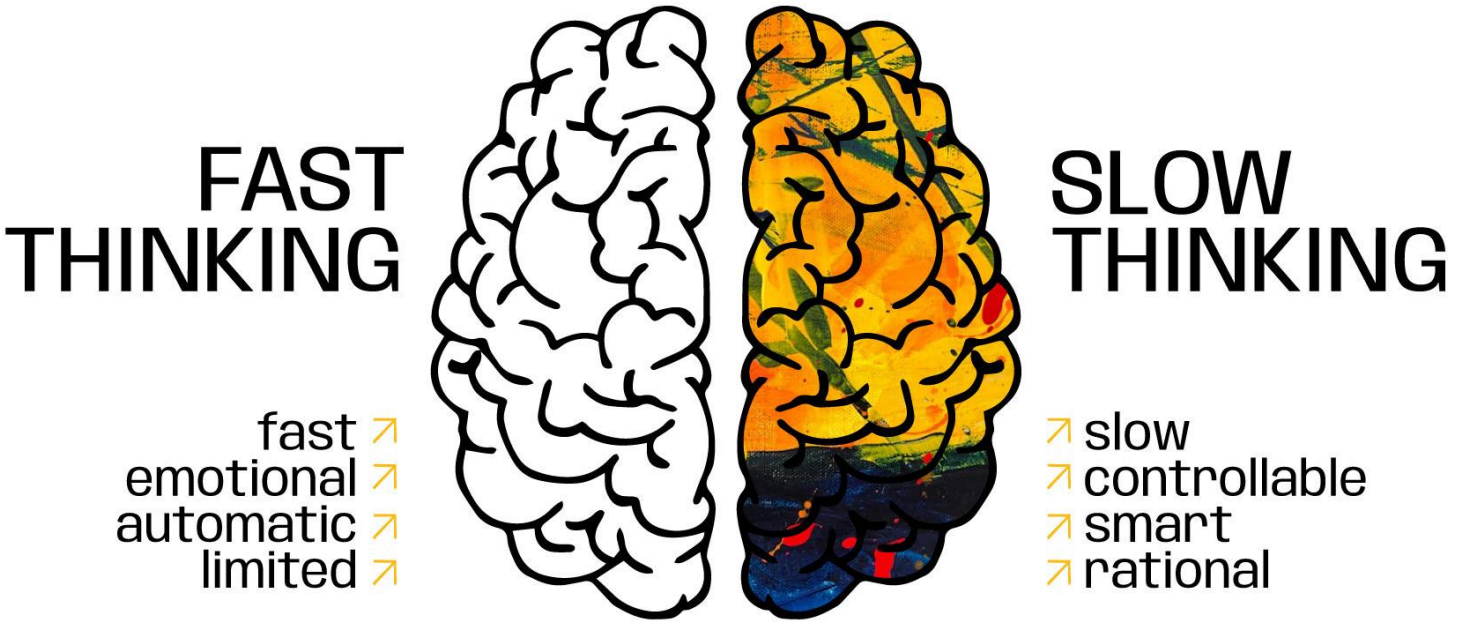
- Test-time Compute: Instead of continuously increasing pre-training budgets, test-time compute allows models to "think longer" during inference.





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 $\boxed{18.75}$

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.

Without modulation

Question 1

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 $\boxed{3}$ **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

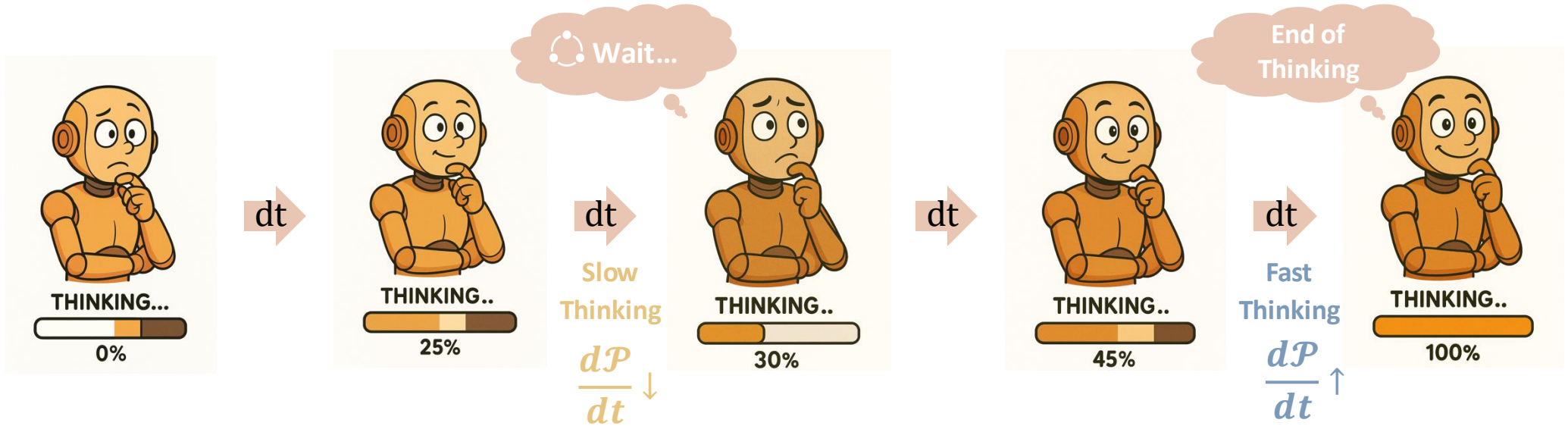
```
Think step by step to answer the following question.
Return the answer at the end of the response after a
separator ####.
```

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

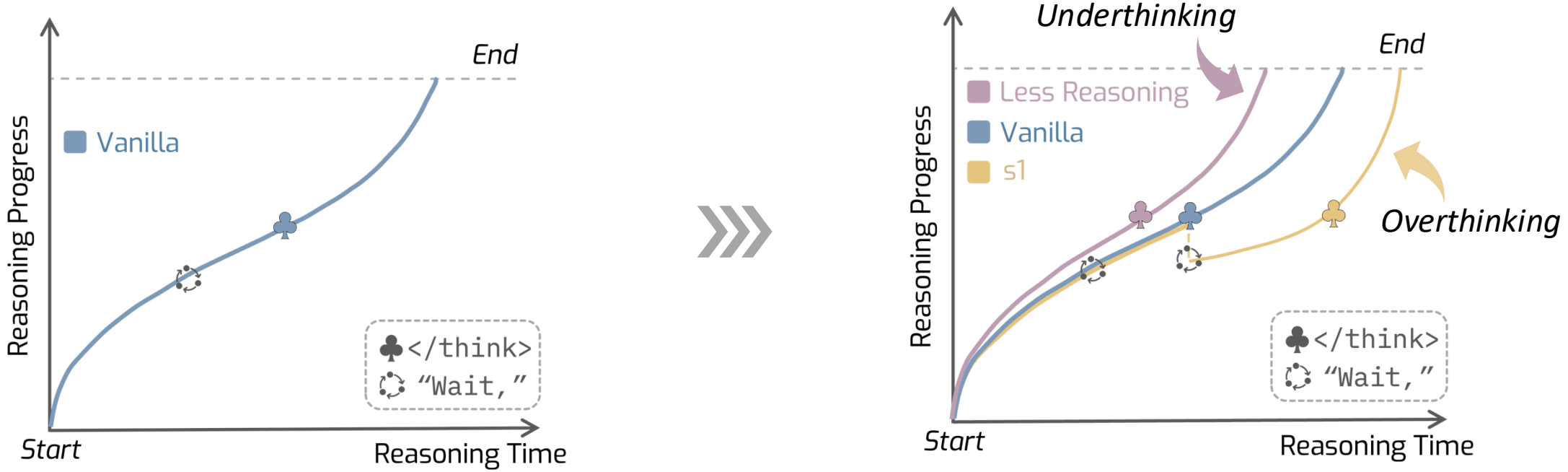
$\frac{d\mathcal{P}}{dt}$: Reasoning Velocity

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

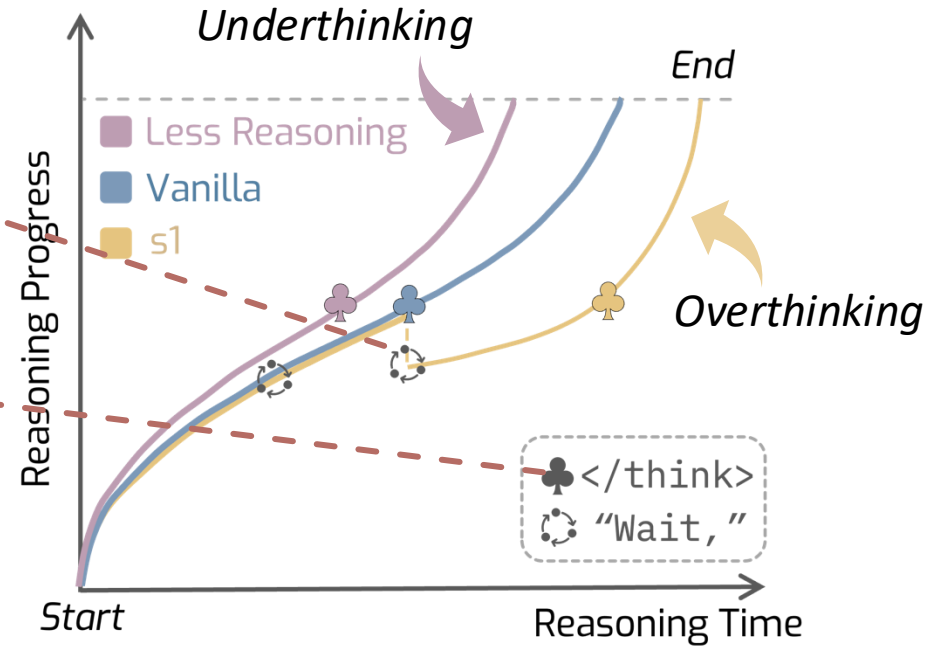
Overthinking Example:

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:

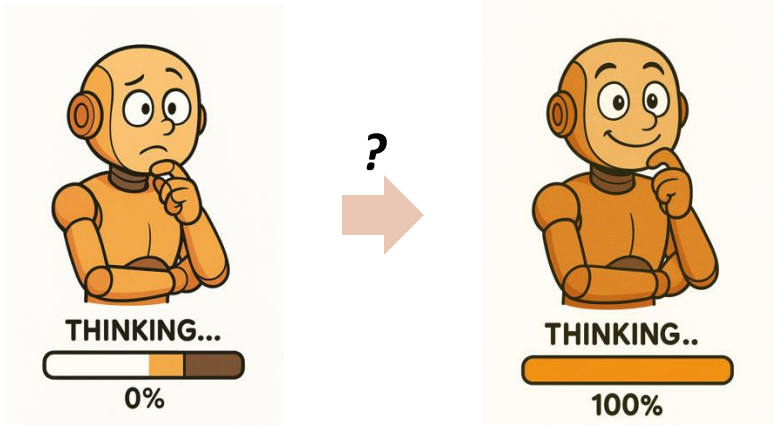
Okay, so the problem given is …… Perhaps I can solve it by …… I can solve …… this problem can be regarded as …… Based on this information, I believe the answer to this problem is ……

(Wrong answer) ✕

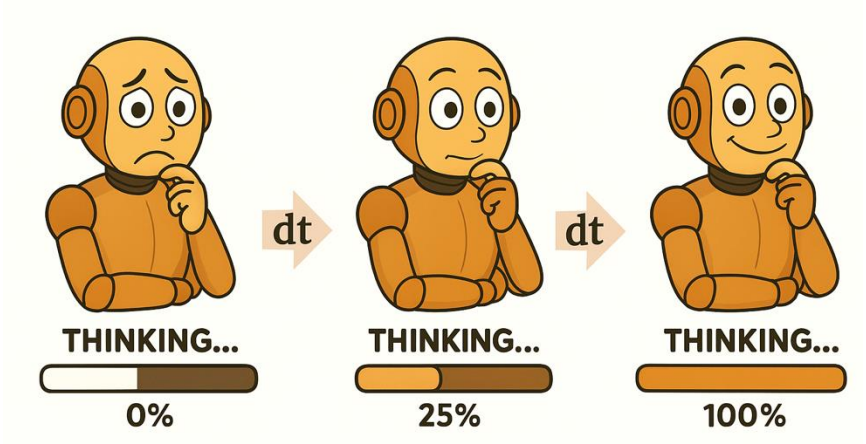
Question 2

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

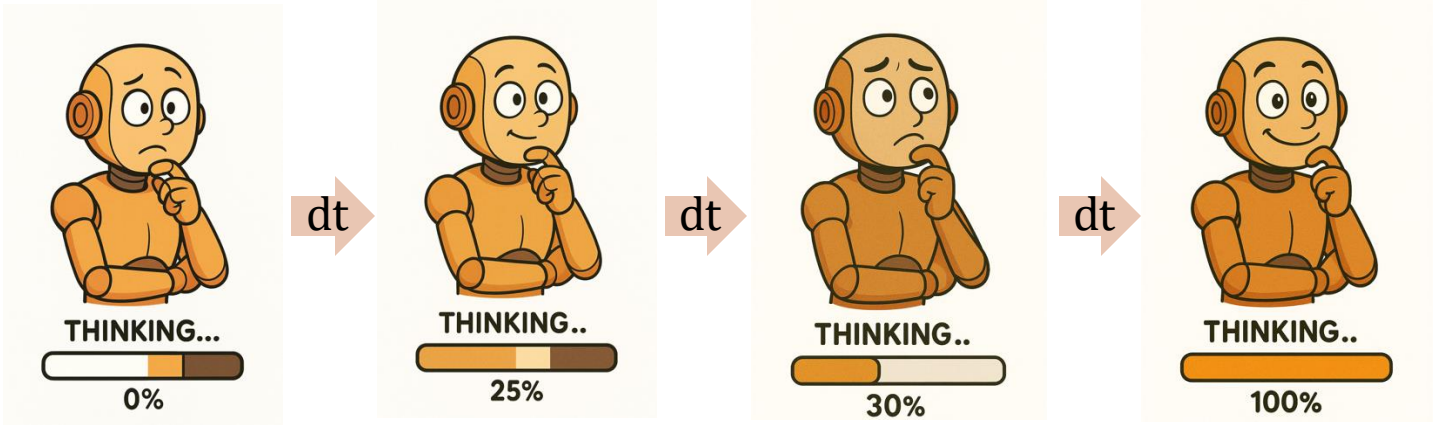
AlphaOne: A Universal View of Test-Time Scaling



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

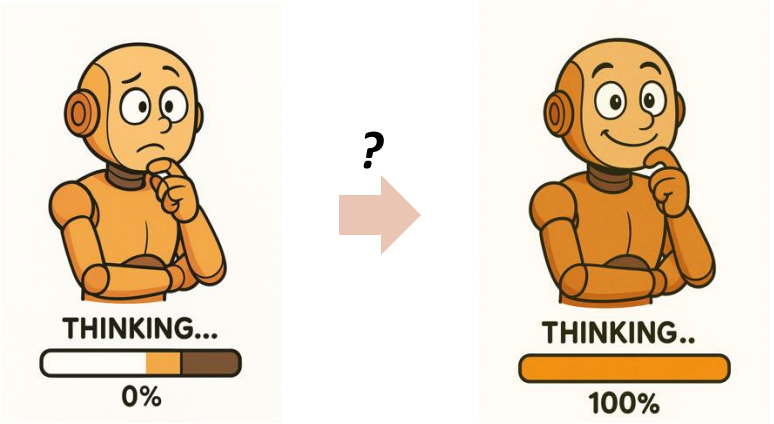


Less Thinking Phase Budget

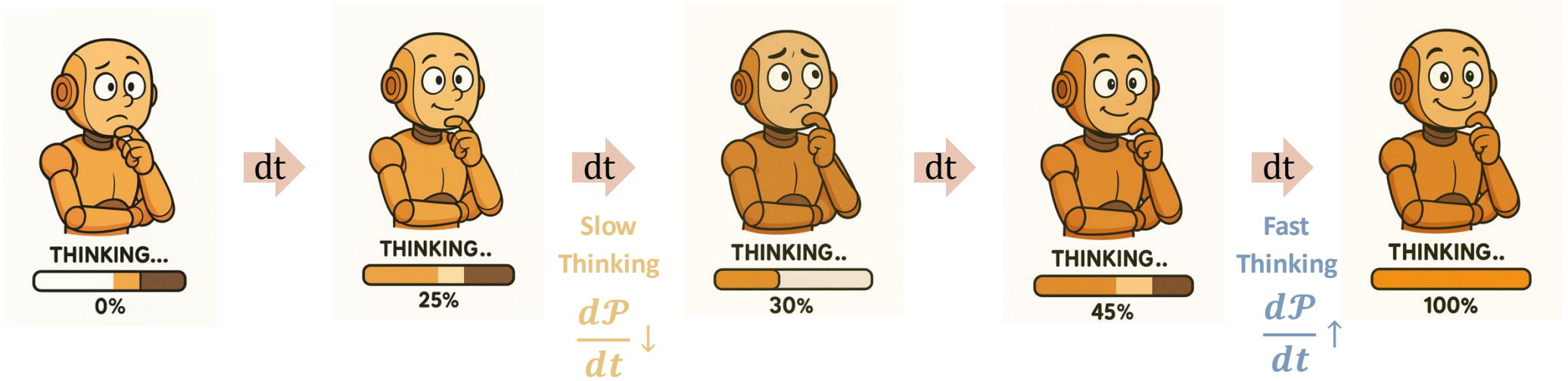


More Thinking Phase Budget

AlphaOne: A Universal View of Test-Time Scaling

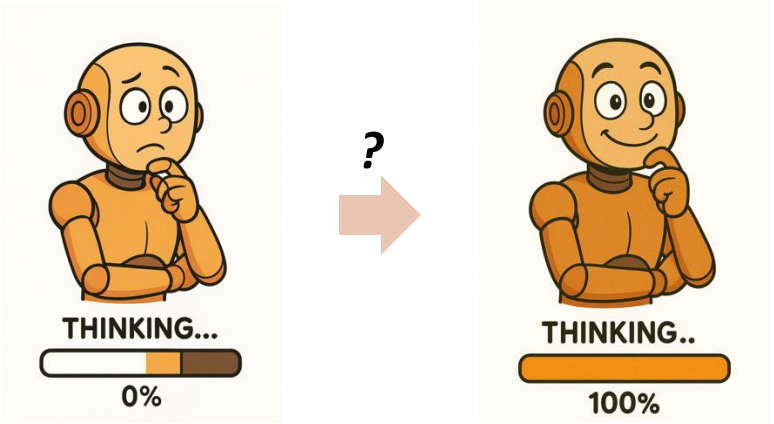


Slow Thinking Scheduling *Modulating slow-fast thinking transitioning arbitrarily.*

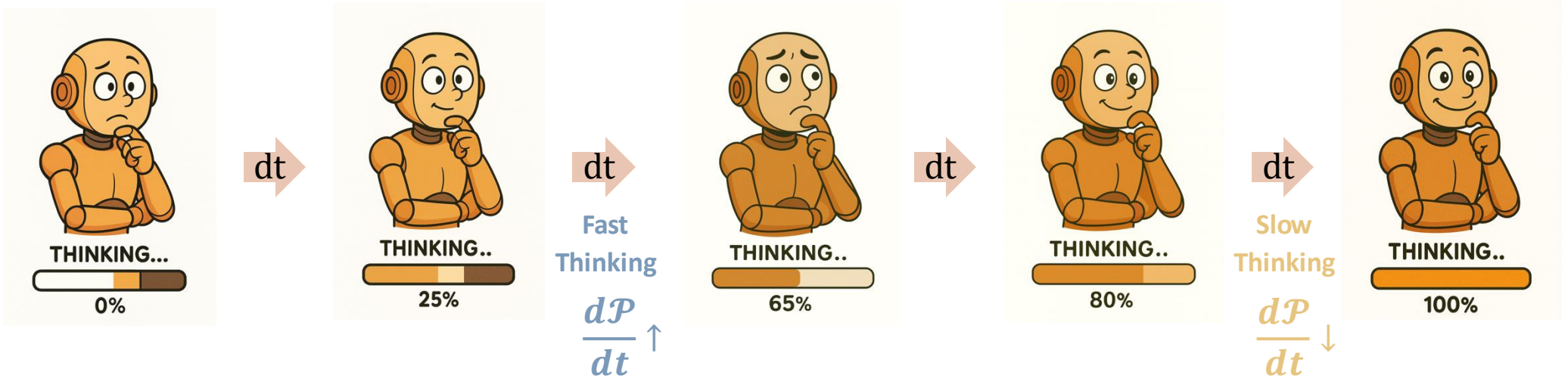


Slow, then Fast

AlphaOne: A Universal View of Test-Time Scaling



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



Fast, then Slow

ALPHAONE: α Moment for Universal Modulation

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

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

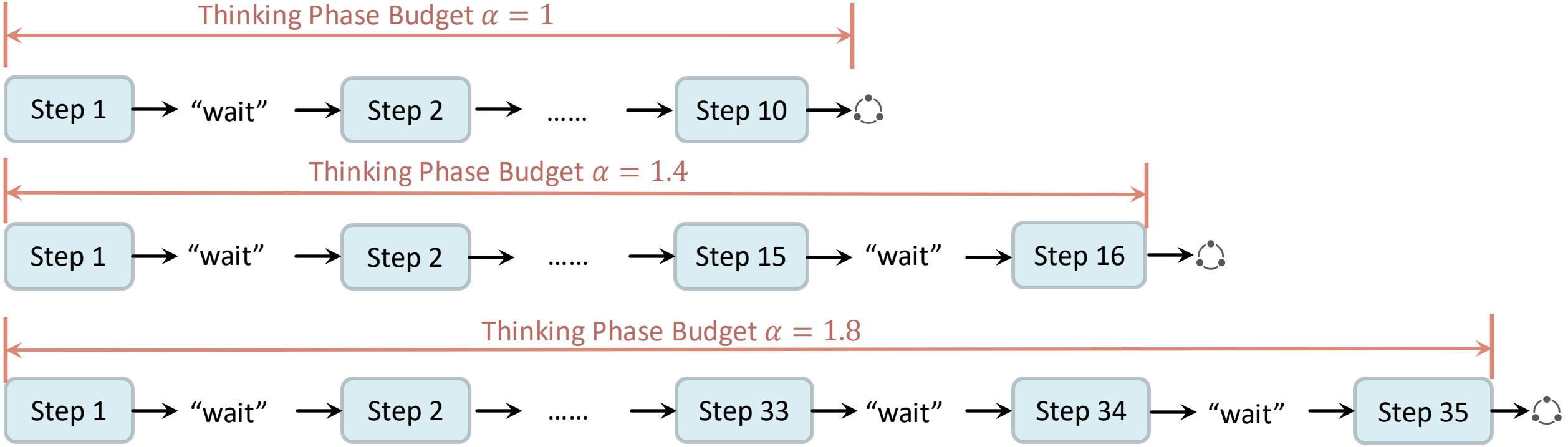
- **Pre α Moment Modulation**

  Slow Thinking Activation: "/n/n" + "wait"

  Stochastic Reasoning Transitioning:

$$\text{Bernoulli}(\boldsymbol{p}_{\texttt{wait}})$$

$$\boldsymbol{p}_{\texttt{wait}} := \mathcal{S}(t), t = 0, 1, \ldots, T_m.$$



AlphaOne: α Moment for Universal Modulation

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

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

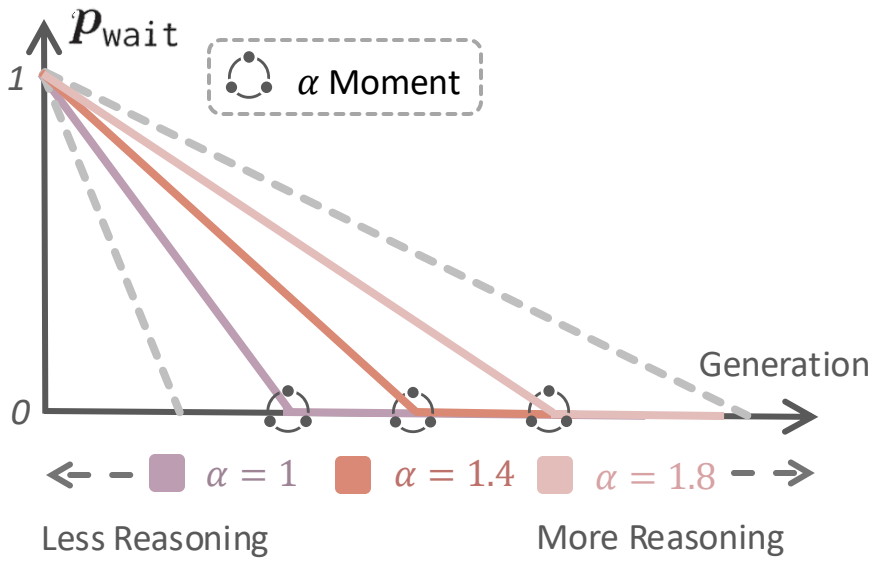
- **Pre α Moment Modulation**

  Slow Thinking Activation: "/n/n" + "wait"

  Stochastic Reasoning Transitioning:

$$\text{Bernoulli}(\boldsymbol{p}_{\texttt{wait}})$$

$$\boldsymbol{p}_{\texttt{wait}} := \mathcal{S}(t), t = 0, 1, \ldots, T_m.$$



ALPHAONE: α Moment for Universal Modulation

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

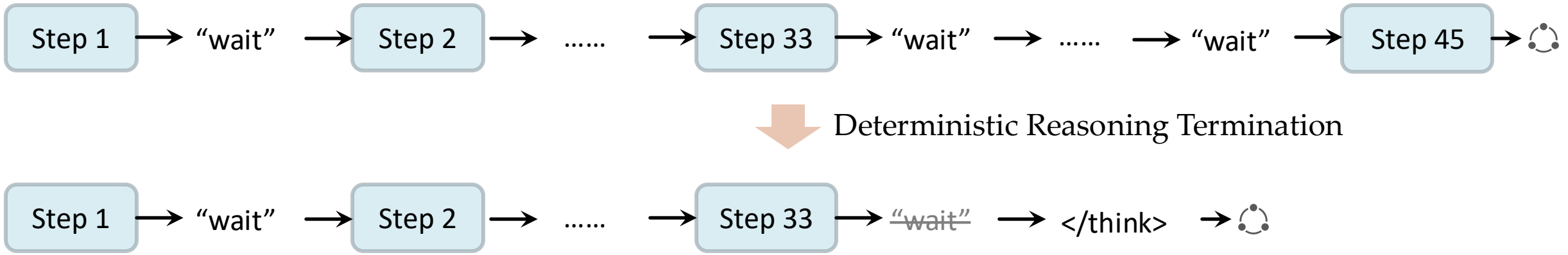
Given the average thinking phase token length (heuristic) $\overline{N}_{\text{think}} > 0$ *and* $\alpha > 0$*, the scaled thinking phase token length becomes* $T_m = \alpha\overline{N}_{\text{think}}$.

- **Post α Moment Modulation**

Deterministic Reasoning Termination:

"wait" $\Longrightarrow$ "</think>"

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



AlphaOne: System Illustration

Question

Let K be the number of sequences A_1, A_2, ... , 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

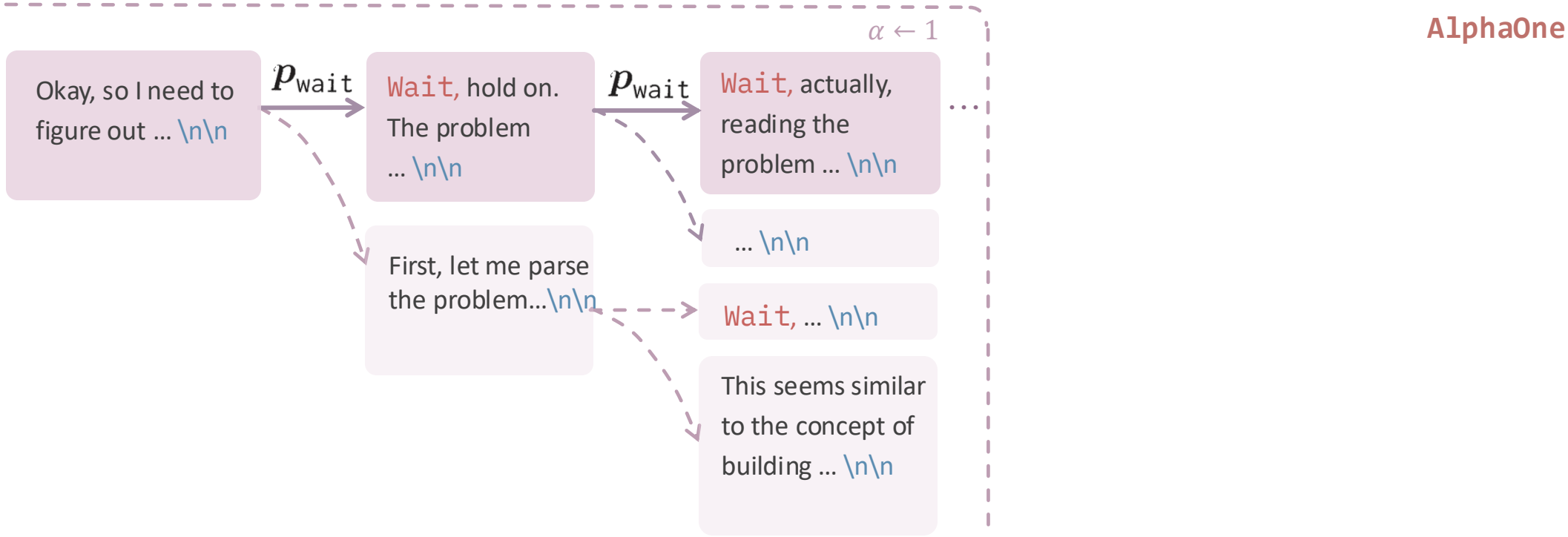
Question

Let K be the number of sequences A_1, A_2, ... , 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

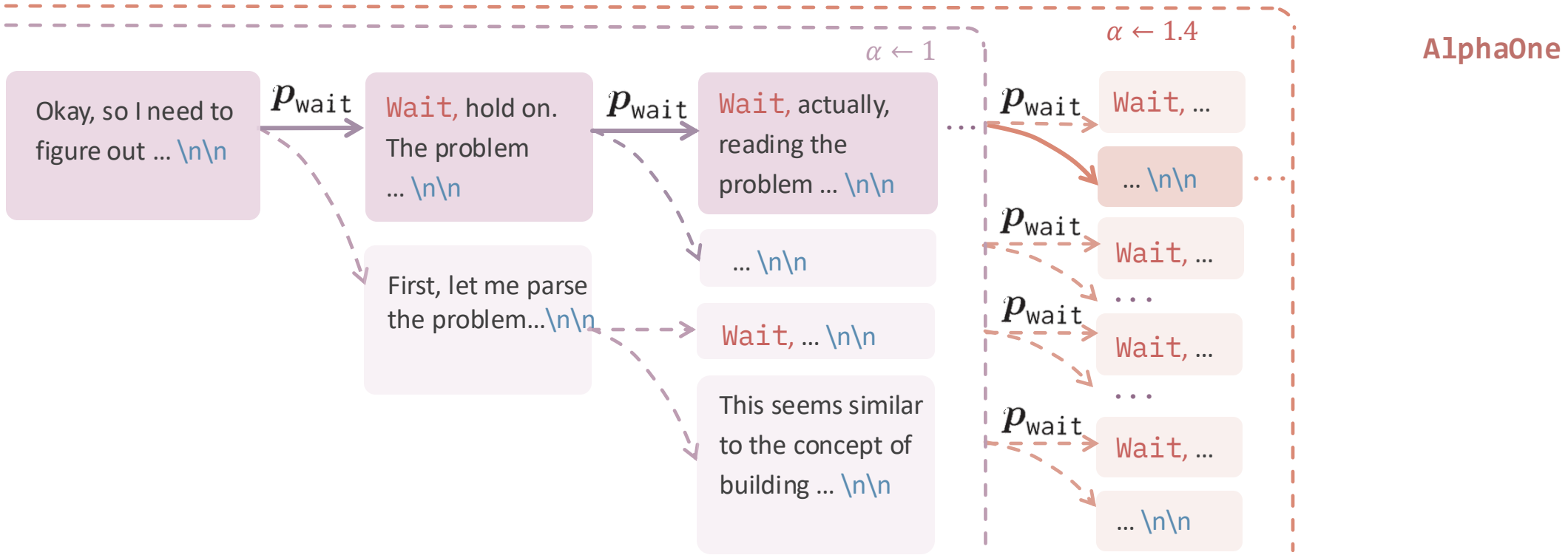
Question

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

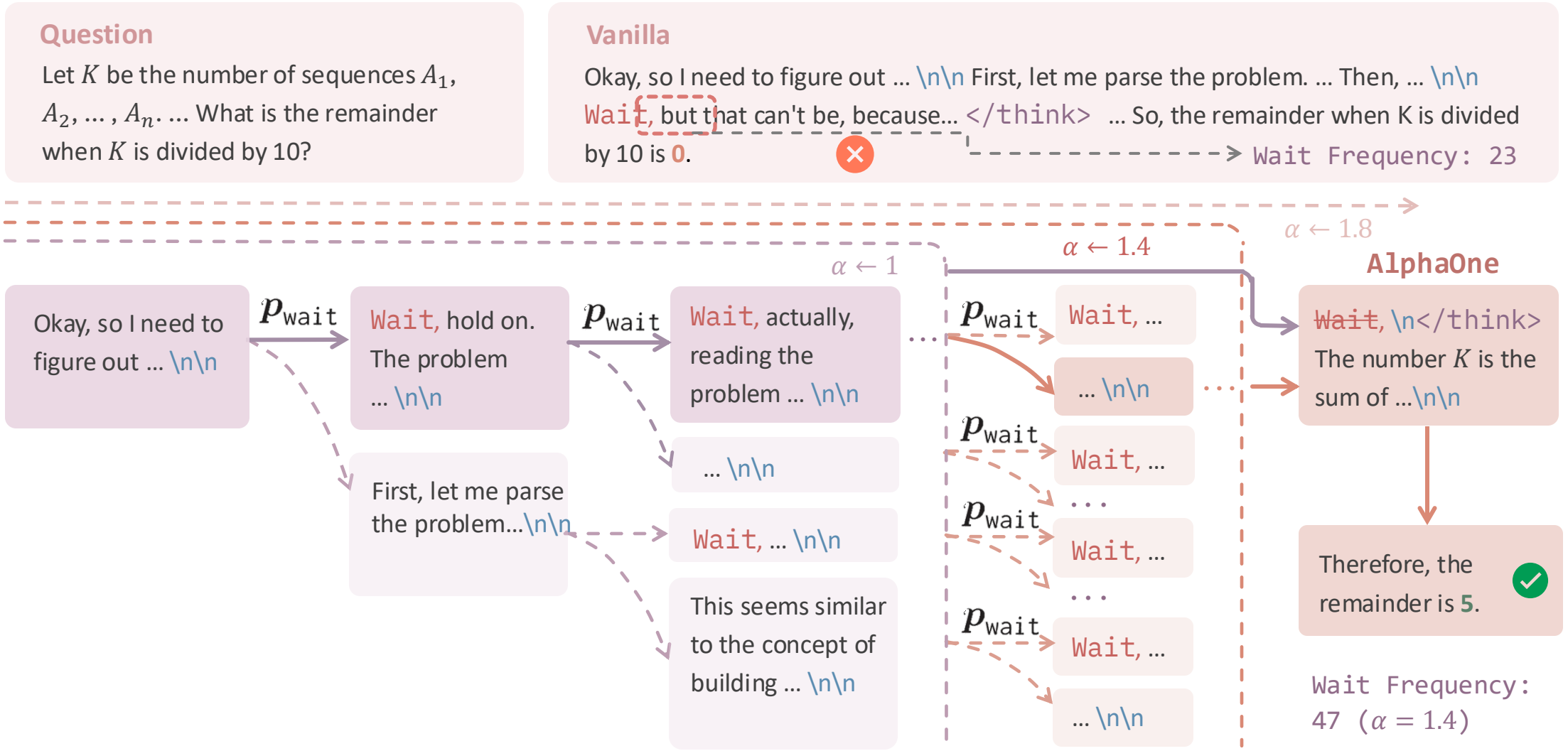
Vanilla

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.

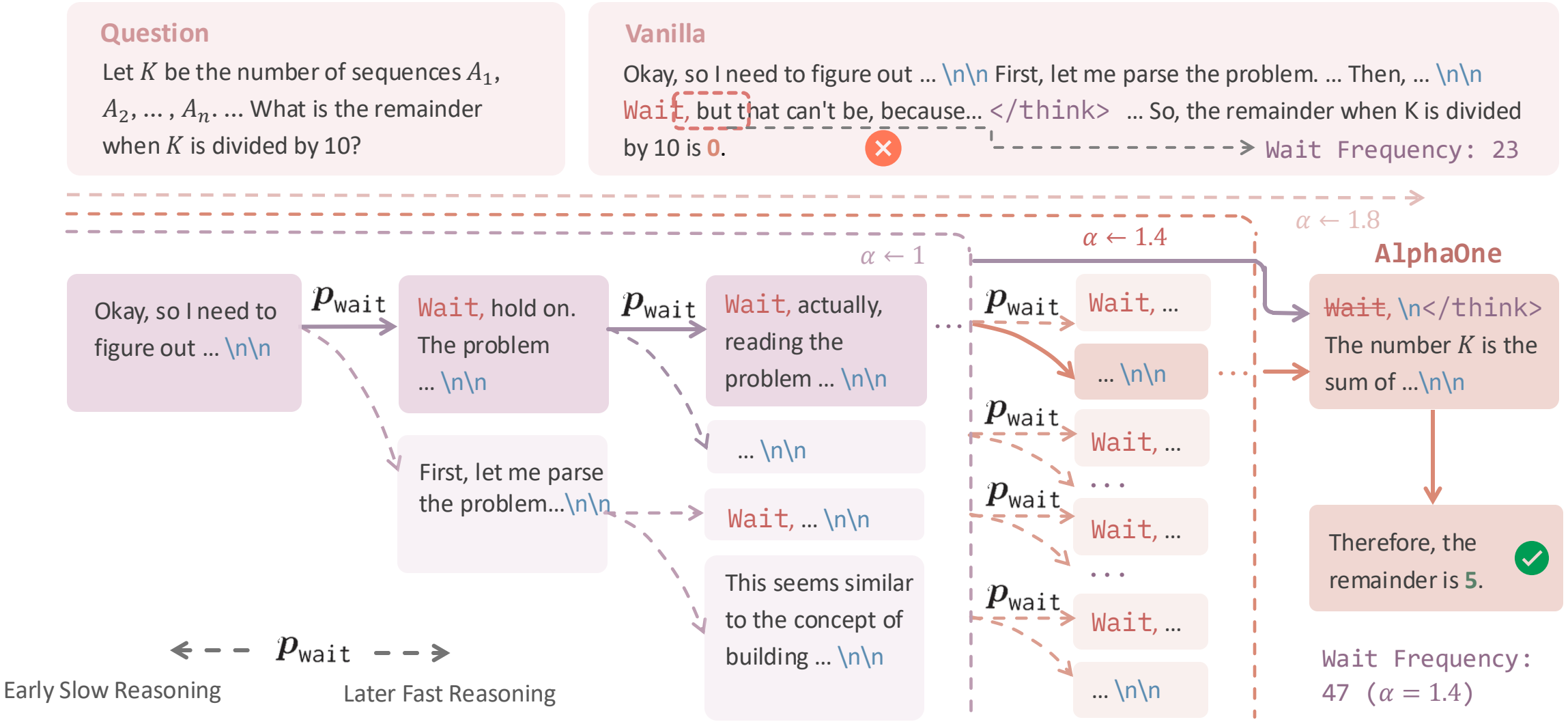
Wait Frequency: 23



AlphaOne: System Illustration



AlphaOne: System Illustration

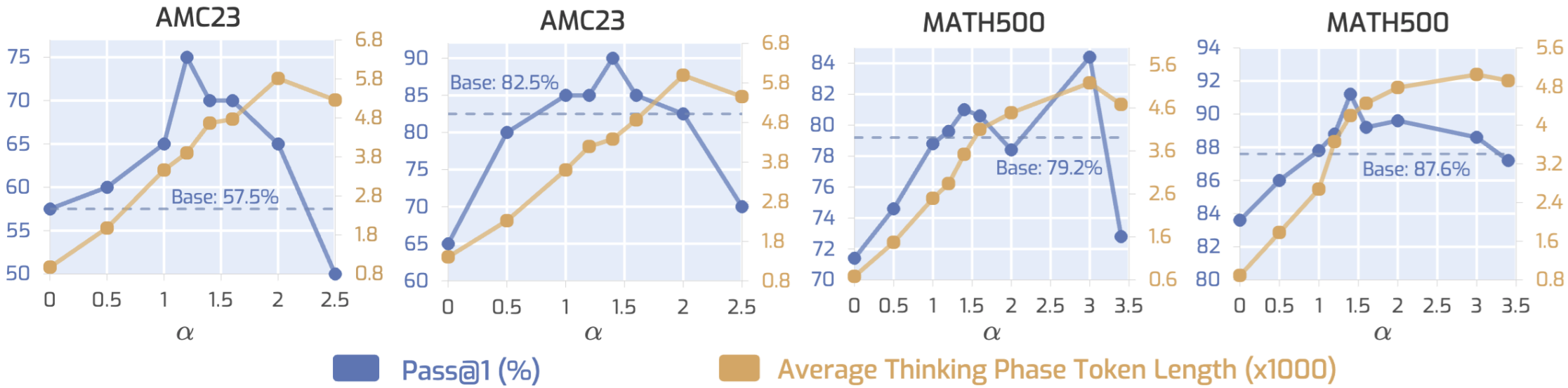


AlphaOne: Systematic Comparison

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

AlphaOne: Can α-moment scale the thinking phase budget?

- **Scaling property of α:** scale α from 0 to the maximum value
  1. α-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 α and the resulting reasoning accuracy

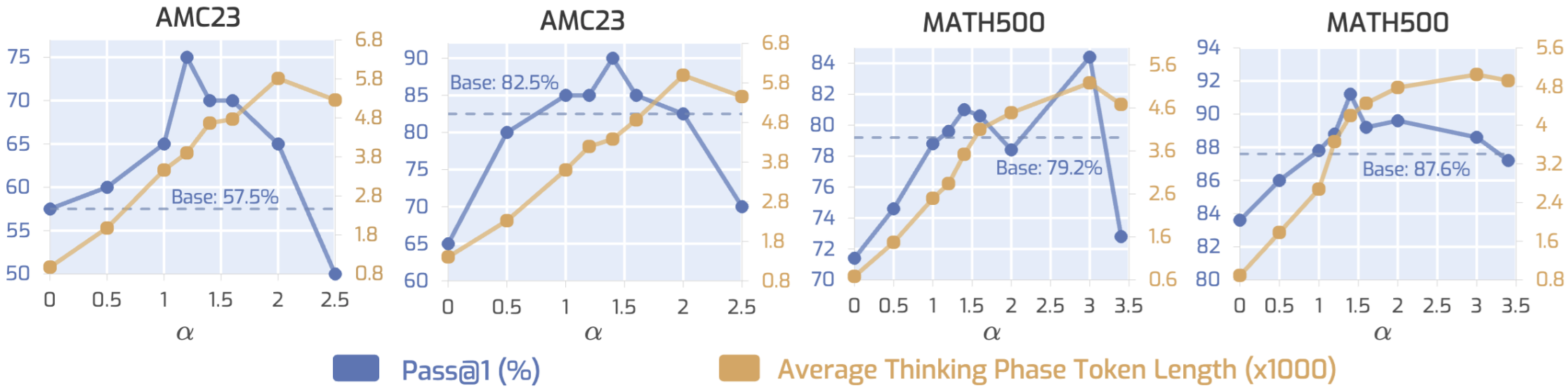


(a) DeepSeek-R1-Distill-Qwen-1.5B (b) DeepSeek-R1-Distill-Qwen-7B (c) DeepSeek-R1-Distill-Qwen-1.5B (d) DeepSeek-R1-Distill-Qwen-7B

ALPHAONE: Can α-moment scale the thinking phase budget?

- **Scaling property of α:** scale α from 0 to the maximum value

  3 The **robustness** of α: Across different values of α, $\alpha 1$ consistently outperforms the base model by a substantial margin

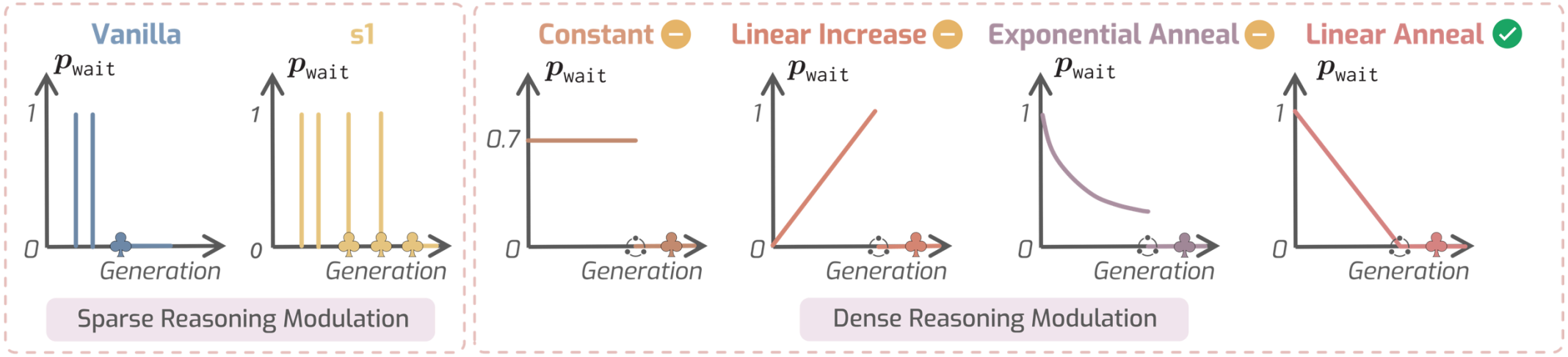


(a) DeepSeek-R1-Distill-Qwen-1.5B (b) DeepSeek-R1-Distill-Qwen-7B (c) DeepSeek-R1-Distill-Qwen-1.5B (d) DeepSeek-R1-Distill-Qwen-7B

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:** $\mathcal{S}(t) := -\frac{1}{T_m} t + 1$

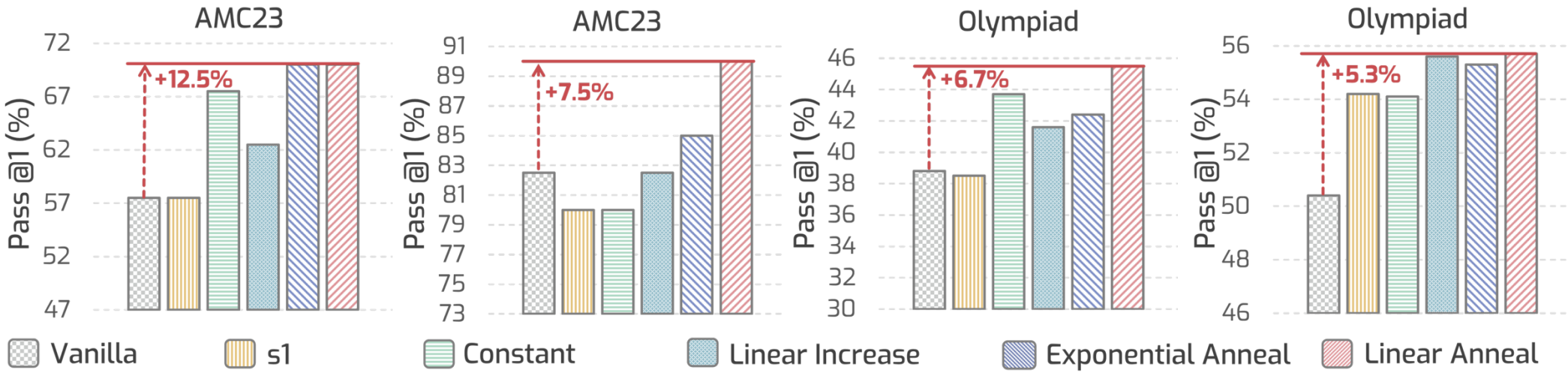
$T_m = \alpha \overline{N}_{\text{think}}$ represent the timestamp of α-moment, $t = \{0,1,\ldots,T_m\}$, and $\gamma > 0$



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



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

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

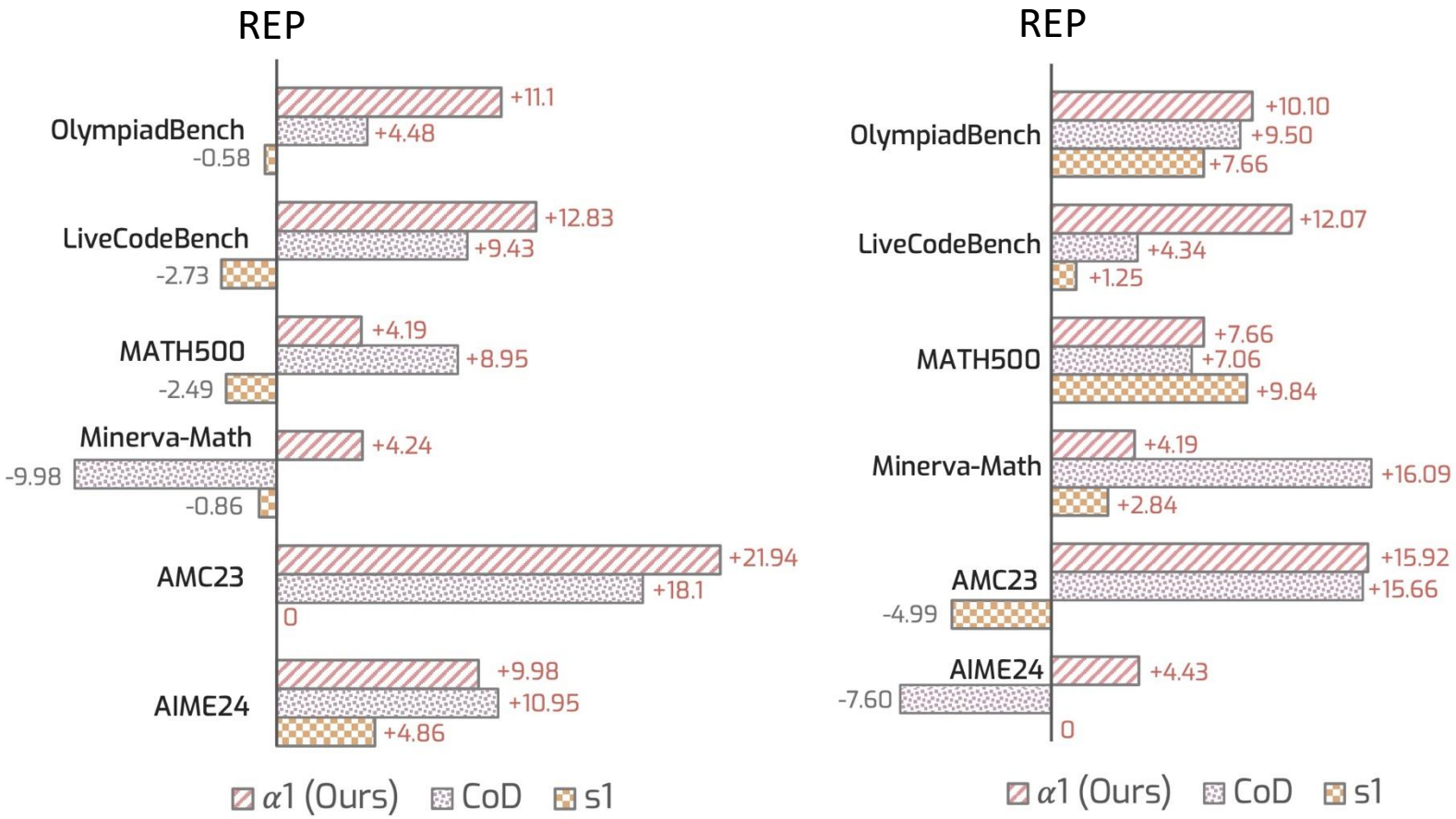
$$\mathcal{F}_{\text{REP}}(\mathcal{A}_{\text{method}}; \mathcal{A}_{\text{base}}, T_{\text{norm}}) = \frac{\mathcal{A}_{\text{method}} - \mathcal{A}_{\text{base}}}{T_{\text{norm}}}$$

Normalized thinking phase token length

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

AlphaOne: Does α1 scale more efficiently?

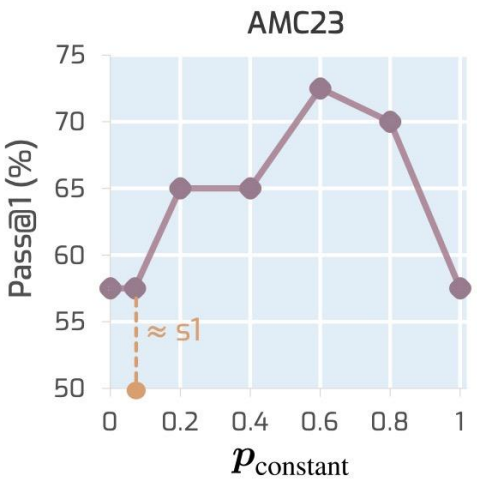
Finding 2 Slow thinking can bring efficient test-time scaling.



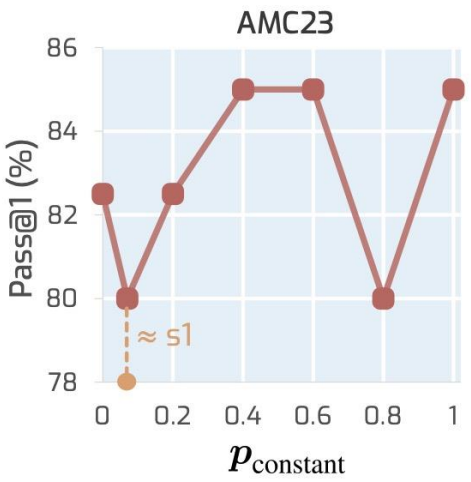
AlphaOne: How frequent should slow thinking transitioning be?

- **Scaling property of "wait" frequency under constant scheduling**

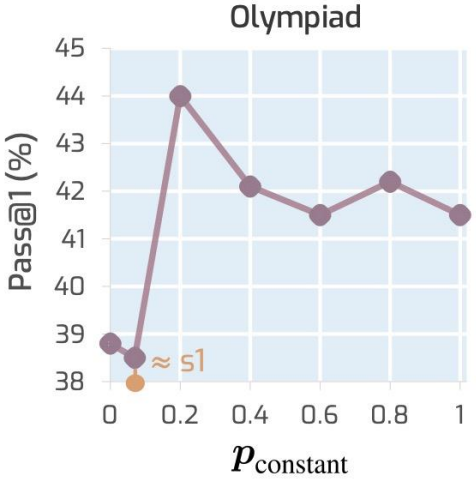
Finding 3 Slow thinking scheduling in high frequency is helpful.



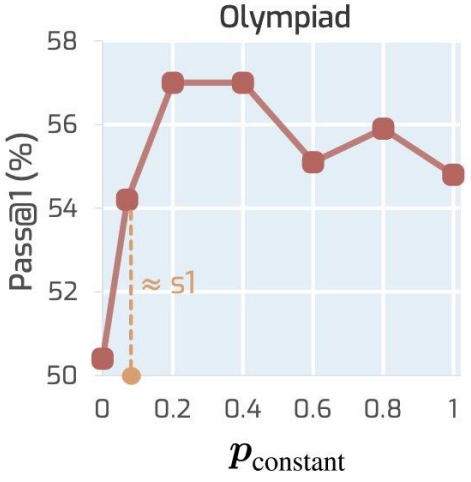
(a) DeepSeek-R1-Distill-Qwen-1.5B



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

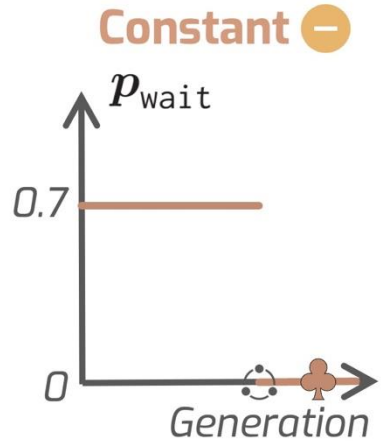


(c) DeepSeek-R1-Distill-Qwen-1.5B



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

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



constant scheduling

ALPHAONE: Is post-α moment modulation necessary?

- Pre-α moment modulation of slow thinking is insufficient

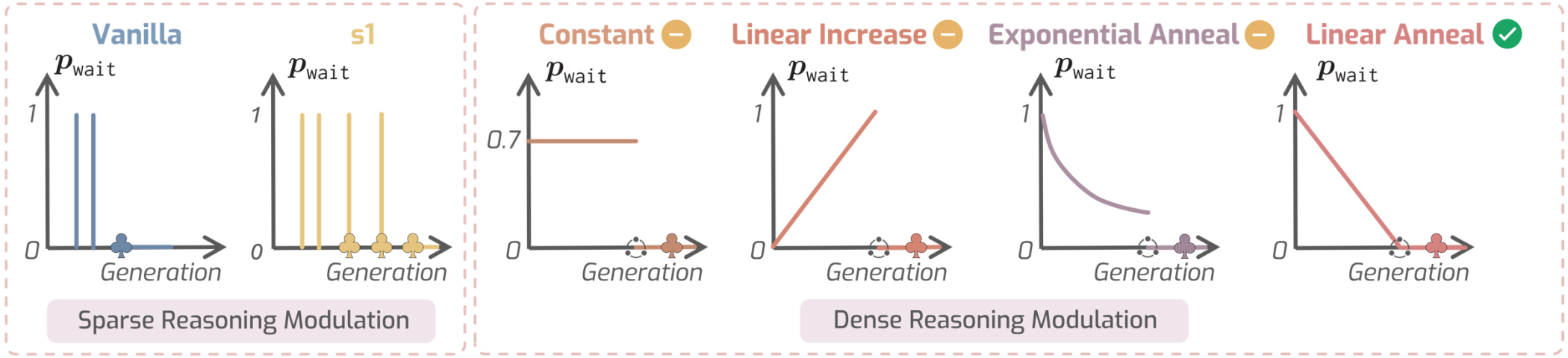
slow thinking inertia

- α1 successfully ends in a fast thinking with post-α moment modulation

Method	Post-α Moment Modulation	AIME24		AMC23	
		P@1	#Tk	P@1	#Tk
DeepSeek-R1-Distill-Qwen-1.5B					
BASE	N/A	23.3	7280	57.5	5339
α1 (Ours)	×	26.7	7929	47.5	6903
α1 (Ours)	✓	**30.0**	**5916**	**70.0**	**4951**
DeepSeek-R1-Distill-Qwen-7B					
BASE	N/A	38.8	**5999**	82.5	4624
α1 (Ours)	×	30.0	7666	75.0	5878
α1 (Ours)	✓	**50.0**	6826	**90.0**	**4397**

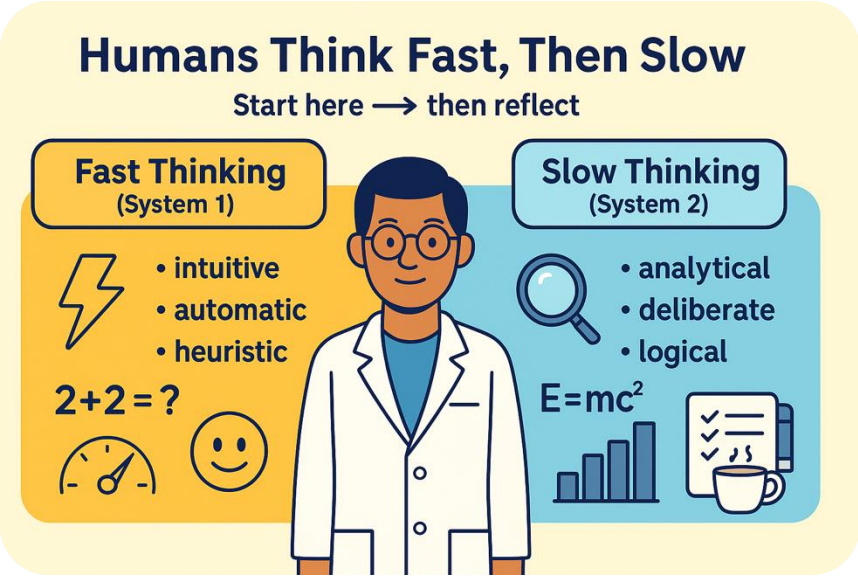
AlphaOne: Beyond the Method

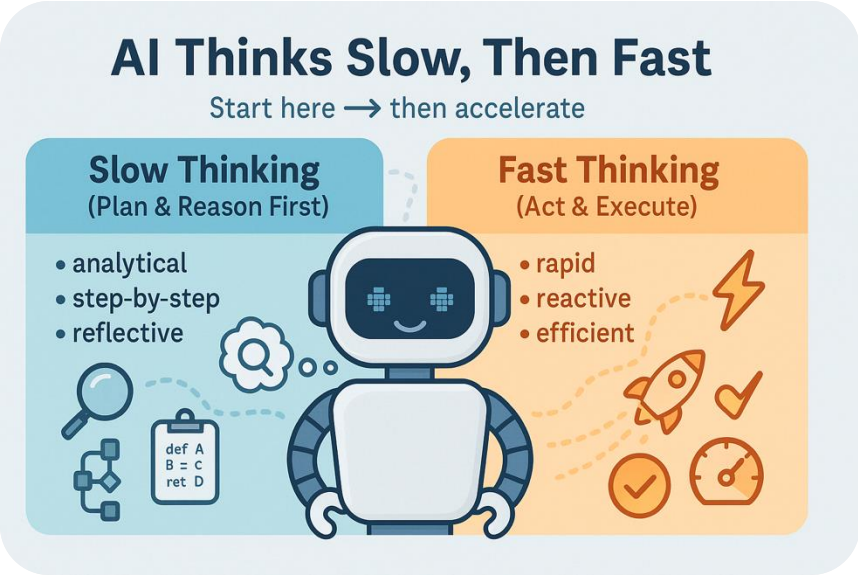
1. **α1 is a universal framework for reasoning modulation:** It unifies and generalizes existing monotonic scaling methods by enabling flexible and dense slow-to-fast reasoning modulation.



AlphaOne: Beyond the Method

2. **α1 challenges prevailing reasoning paradigms:** The reasoning models fail to find the optimal scheduling for switching fast-slow reasoning without modulation.
3. **α1 utilizes a simple method to motivate an eternal problem:** The optimal test-time reasoning modulation is still unknown.





The Future of AlphaOne

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 AlphaOne

α1-style reasoning modulation with RL: Utilize reinforcement learning to learn an adaptive reasoning modulation in α1 fashion. For instance, α-moment can be adaptively sampled from a subnetwork, and the reasoning scheduling strategy can be adaptively selected when facing different problems.

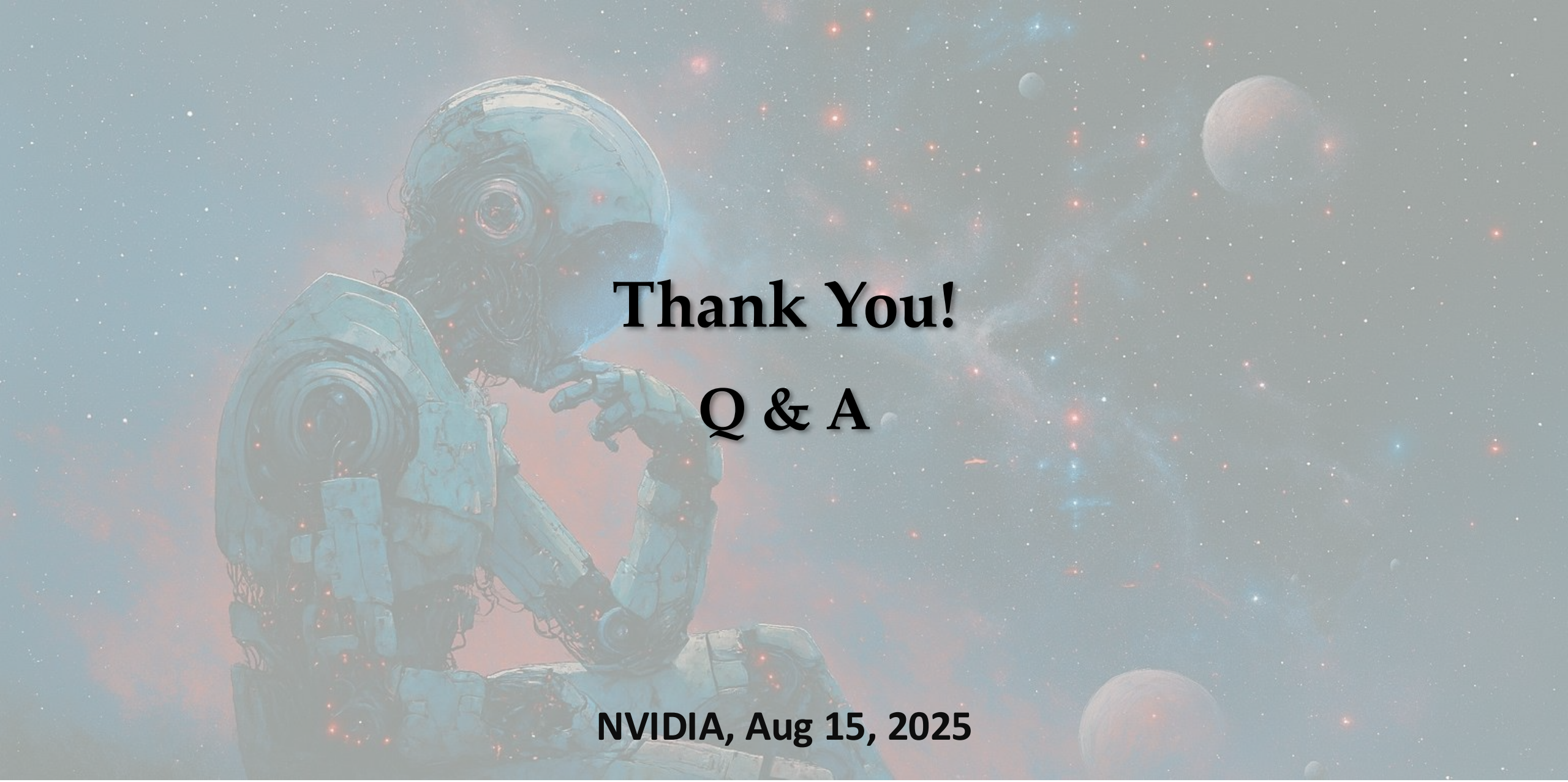
α1 as a sampling approach: utilize α1 to modulate the reasoning progress when sampling rollouts. This brings more reasoning pattern diversity of classical methods like GRPO.

The Future of AlphaOne

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.



Thank You!

Q & A

NVIDIA, Aug 15, 2025

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Berkeley