Think Only When You Need with Large Hybrid-Reasoning Models

Lingjie Jiang[*1,2], Xun Wu[*1], Shaohan Huang[1], Qingxiu Dong[1,2], Zewen Chi[1], Li Dong[1],
Xingxing Zhang[1], Tengchao Lv[1], Lei Cui[1], Furu Wei[1,†]
Microsoft Research[1], Peking University[2]

* Equal contribuition.

Motivation

- Large Reasoning Models (LRMs) have shown strong improvements in reasoning tasks by generating **extended thinking traces**.
- However, LRMs tend to **overthink**, applying complex reasoning even to **simple queries**, such as a single word.

Question 1: Let a, b, c be real numbers such that $|ax^2 + bx + c| \le 1$for all $0 \le x \le 1$. Find the largest possible value of $|a| + |b| + |c|$. Let's think step by step and output the final answer within boxed{}. Ground Truth: 17

Qwen2.5-7B-Instruct: To find the largest possible value of $|a| + |b| + |c|$ given that $|ax^2 + bx + c| \le 1$ for all $0 \le x \le 1$, we need to analyze the constraints imposed by this inequality at specific points.
...
All conditions are satisfied. Now, calculate $|a| + |b| + |c|$: $|a| + |b| + |c| = |2| + |-2| + |1| = 2 + 2 + 1 = 5$ Thus, the largest possible value of $|a| + |b| + |c|$ is $\boxed{5}$. ✗

DeepSeek-R1-Distill-Qwen-7B: `<think>` Okay, so I have this problem where I need to find the largest possible value of |a| + |b| + |c| given that $|ax^2 + bx + c| \le 1$ for all x between 0 and 1. Hmm, let me think about how to approach this. ... `</think>`
The maximum value of |a| + |b| + |c| is achieved when the quadratic function touches 1 at x=0 and x=1, and -1 at x=1/2.
...
For $0 \le x \le 1$, $0 \le (x - \frac{1}{2})^2 \le \frac{1}{4}$, so $-1 \le f(x) \le 1$. Therefore, the largest possible value of $|a| + |b| + |c|$ is $\boxed{17}$ ✓

Question 2: Can you help me please?

Qwen2.5-7B-Instruct: Of course! I'd be happy to help you. Could you please provide some more information about what you need assistance with? ✓

DeepSeek-R1-Distill-Qwen-7B: `<think>` Okay, the user just said "Can you help me please?" I need to respond in a friendly and helpful way. I should acknowledge their request, show willingness to assist, and ask them to specify what they need help with. Keeping it open-ended will encourage them to provide more details.
Hmm, maybe start with a positive affirmation. Something like, "Yes, I can help! What do you need assistance with today?" That sounds good. It's polite and inviting them to share more about their request.
I should make sure my response is clear and concise. No need for complicated words. Just a straightforward offer to help. Let me put that together. `</think>`
Of course! I'm here to help. What do you need assistance with today? ✗

Fig.1 A case of LLM and LRM.

How to build a hybrid thinking system that can achieve an optimal balance between system 2 reasoning and system 1 ability?

Method – Large Hybrid-Reasoning Models

We introduce **Large Hybrid-Reasoning Models (LHRMs)** that can dynamically choose between:

- **Thinking Mode**: engage in multi-step reasoning for complex queries
- **No-Thinking Mode**: respond directly to simple questions

Question 1: Let a, b, c be real numbers such that $|ax^2 + bx + c| \le 1$for all $0 \le x \le 1$. Find the largest possible value of $|a| + |b| + |c|$. Let's think step by step and output the final answer within boxed{}. Ground Truth: 17

LHRMs-7B: <think> Okay, so I need to find the largest possible value of |a| + |b| + |c| given that the quadratic function $|ax^2 + bx + c|$ is bounded by 1 in absolute value for all x between 0 and 1. Hmm, let's start by understanding the problem. ... </think>
Given the problem of finding the largest possible value of $|a| + |b| + |c|$ under the constraint that $|ax^2 + bx + c| \le 1$ for all $0 \le x \le 1$, we start by considering the Chebyshev polynomials and their properties.
...
After verifying that this polynomial indeed satisfies the given constraint and considering other potential polynomials, we conclude that the maximum value of $|a| + |b| + |c|$ is achieved by the polynomial $8x^2 - 8x + 1$. The largest possible value of $|a| + |b| + |c|$ is $\boxed{17}$. ✓

Question 2: Can you help me please?

LHRMs-7B: <no_think> Of course! What do you need help with?</no_think> ✓

Fig.2 A case of LHRMs-7B

Model Training – Stage II: Hybrid Group Policy Optimization (HGPO)

- **Goal:** Learn a policy to adaptively select the appropriate reasoning mode (Thinking or No-Thinking) per query, while enhancing helpfulness and reasoning quality.

- **Method**
  - For each query, sample responses under both modes
  - Use a reward model to evaluate and compare responses
  - Combine intra-group (within-mode) and inter-group (between-mode) rewards
  - Learn to prefer the most effective reasoning mode per query

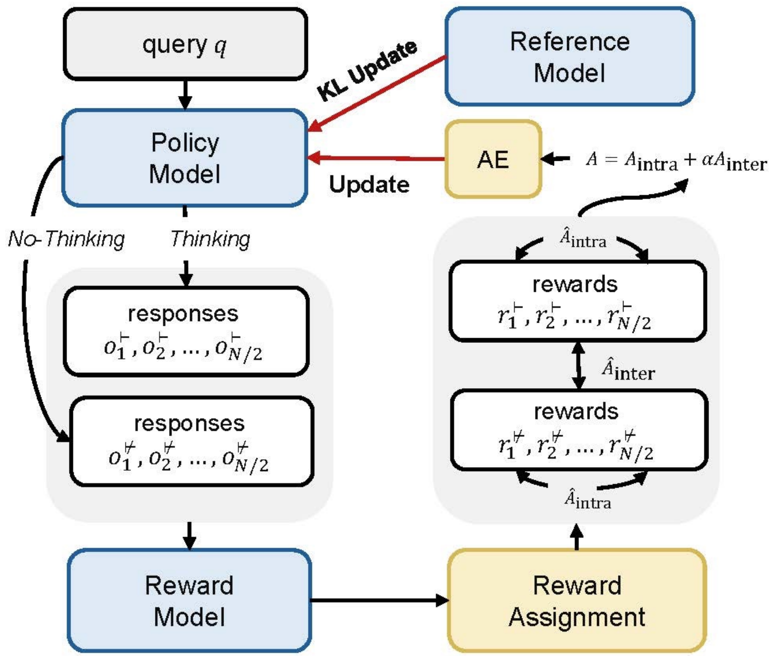
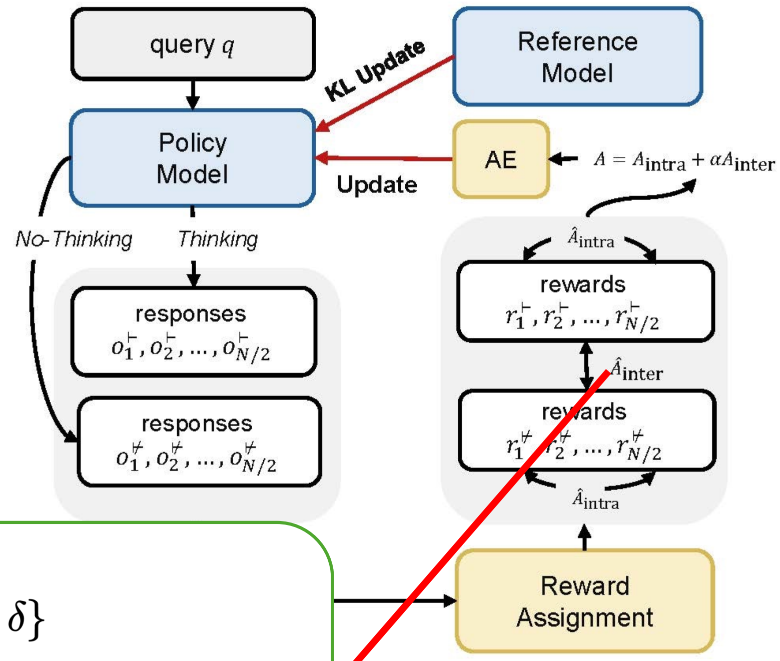


Fig.3 Diagram of the HGPO Algorithm

Model Training – Stage II: Hybrid Group Policy Optimization (HGPO)

- **Goal:** Learn a policy to adaptively select the appropriate reasoning mode (Thinking or No-Thinking) per query, while enhancing helpfulness and reasoning quality.

- **Method**
  - For each query, sample responses under both modes
  - Use a reward model to evaluate and compare responses



the HGPO Algorithm

$$r_{\text{inter}}(o_i^m) = \begin{cases} 1, & \text{if } m = \underset{m' \in \{\text{T,NT}\}}{\operatorname{argmax}} \{\bar{\mathcal{R}}^{\text{T}}, \bar{\mathcal{R}}^{\text{NT}} + \delta\} \\ 0, & \text{otherwise} \end{cases}$$

Margin δ is a hyperparameter that controls the **trade-off between Thinking and No-Thinking modes** during reward assignment

Evaluating Hybrid Thinking Capability

- We propose a new metric, Hybrid Accuracy (H_{Acc}), which measures the model's ability to correctly choose between Thinking and No-Thinking modes.

- **Definition**

$$H_{acc} = \frac{1}{K} \sum_{i=1}^{K} \mathbb{1}[Equal(m_{gt}, m_p)]$$

  - Generate responses in **both modes**
  - Score them using a **reward model**
  - Identify the **better-performing mode** → m_{gt}
  - Get the model's **selected mode** → m_p
  - Compute the percentage of queries where $m_{gt} = m_p$

Experiments — Main Results

- LHRMs consistently outperform baseline LLMs, LRMs, and hybrid variants trained only with HFT or HFT+DPO/RFT on both reasoning (math, code) and general capabilities (AlpacaEval, Arena-Hard)
- LHRMs achieve the highest Hybrid Accuracy (H_{Acc}), significantly outperforming other hybrid baselines (HFT, HFT-DPO/RFT), proving that HGPO effectively teaches mode selection

Methods	Type	MATH				Code			General		$\mathcal{H}_{\text{acc}}$	Avg.
		MATH500	AIME24	AMC23	Olympiad	LiveCode	MBPP	MBPP+	Alpaca	Arena		
1.5B size model												
Qwen2.5-Math-1.5B	LLMs	42.4	3.3	22.5	16.7	0.4	16.1	14.3	0.1	1.8	-	13.1
Qwen2.5-1.5B-Instruct	LLMs	51.0	3.3	52.8	38.7	2.2	60.1	51.9	8.8	1.1	-	30.0
Qwen2.5-Math-1.5B-Instruct	LLMs	72.0	6.7	60.0	38.1	3.7	26.7	23.8	2.8	4.7	-	26.5
DeepSeek-R1-Distill-Qwen-1.5B	LRMs	83.9	28.9	62.9	43.3	16.8	54.2	46.3	5.6	2.7	-	38.3
HFT-1.5B	Hybrid	**87.8**	32.7	**75.0**	48.9	15.7	54.8	47.4	13.1	6.9	41.4	42.5
HFT-RFT-1.5B	Hybrid	82.2	22.0	67.5	44.1	14.2	49.7	42.6	13.6	8.5	48.1	38.3
HFT-DPO-1.5B	Hybrid	86.8	32.6	**75.0**	48.7	**17.2**	50.5	42.6	13.3	6.9	45.8	41.5
LHRMs-1.5B	Hybrid	**87.8**	**35.3**	**75.0**	**50.4**	**17.2**	**61.1**	**54.0**	**16.9**	**10.4**	**54.4**	**45.3**
7B size model												
Qwen2.5-Math-7B	LLMs	57.0	13.3	22.5	21.8	6.0	31.5	27.3	2.0	7.0	-	20.9
Qwen2.5-7B-Instruct	LLMs	77.0	13.3	52.8	29.1	14.6	79.9	67.5	**36.2**	25.8	-	44.0
Qwen2.5-Math-7B-Instruct	LLMs	82.4	10.0	62.5	41.6	2.6	40.2	34.7	3.8	10.0	-	32.0
DeepSeek-R1-Distill-Qwen-7B	LRMs	92.8	55.5	91.5	58.1	37.6	74.3	64.3	19.1	17.9	-	56.8
HFT-7B	Hybrid	93.6	56.7	**95.0**	58.5	34.7	70.6	59.8	23.7	14.0	34.2	56.4
HFT-RFT-7B	Hybrid	87.8	55.3	82.5	55.0	35.8	81.0	68.8	28.1	14.0	49.7	56.6
HFT-DPO-7B	Hybrid	**93.8**	58.7	92.5	60.6	**38.8**	80.1	68.3	23.3	13.0	37.1	58.9
LHRMs-7B	Hybrid	**93.8**	**66.7**	**95.0**	**61.2**	**38.8**	**81.5**	**69.6**	35.0	**26.0**	**71.9**	**63.1**

Tab.4 Performance comparison between LHRMS

Experiments — Ablation study

Effect of Advantage Estimators

- HGPO supports multiple advantage estimators: **REINFORCE++**, **GRPO**, and **RLOO**
- These estimators all achieve **competitive performance**, indicating that HGPO is **robust to the choice of advantage estimator**

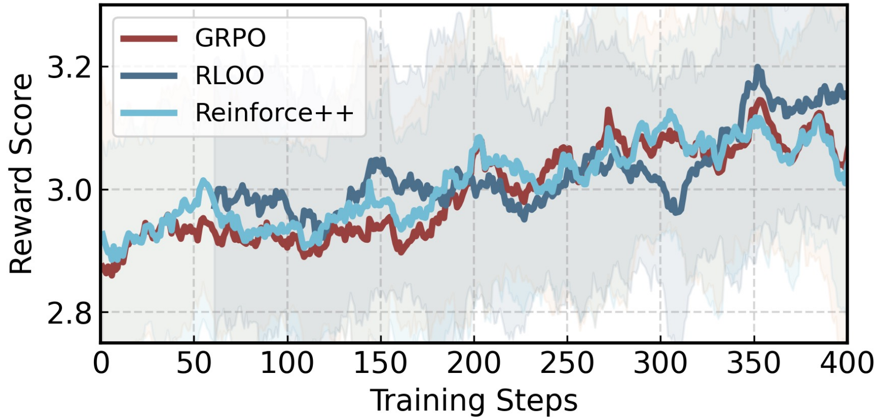


Fig.3 Ablation study on advantage estimators.

Effect of Margin δ in HGPO

- A larger margin δ encourages more use of No-Thinking, shifting the model toward faster responses

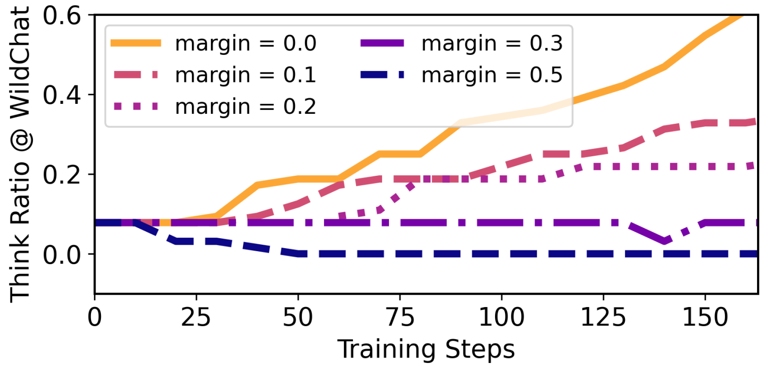


Fig.4 Ablation study on margin δ

Experiments — Thinking Ratio Study

- **Within-domain:** LHRMs learn to reduce unnecessary thinking on easier problems, showing adaptive reasoning compared to HFT baselines.
- **Across-domain:** The model automatically increases reasoning in unseen domains (e.g., code), demonstrating strong generalization and transferability.

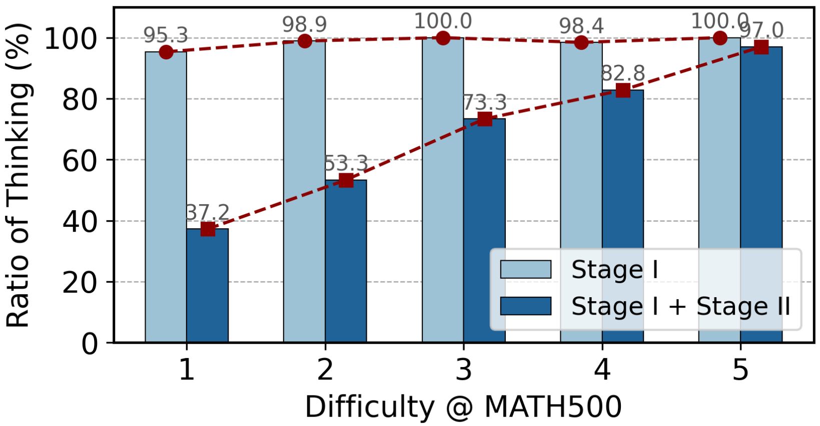


Fig.3 Within-domain (Math)

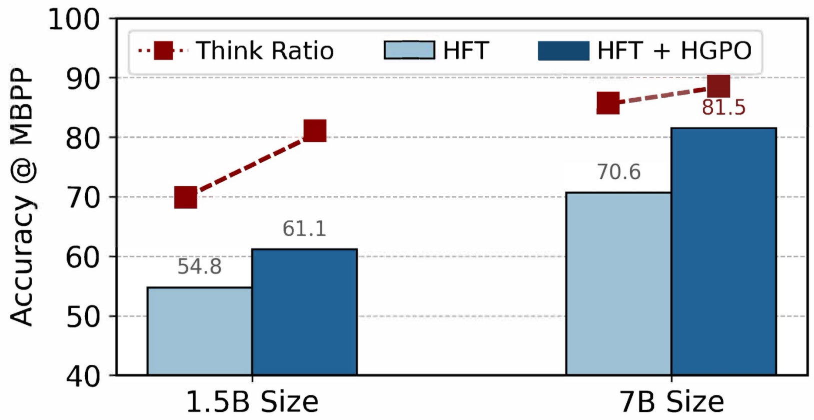


Fig.4 Cross-domain (Code)

Experiments — Efficiency Analysis via Adaptive Thinking

- Compared to always-thinking HFT counterparts, LHRMs-1.5B and LHRMs-7B reduce average thinking tokens by **~23% and ~25% respectively**, while **maintaining or slightly improving** overall performance.

Model	MATH500	AIME24	AMC23	Olympiad	LiveCodeBench	MBPP	MBPP++	AlpacaEval2.0	ArenaHard2.0	Avg.
1.5B Size Models										
DeepSeek-R1-Distill-Qwen-1.5B	83.9 (4059)	28.9 (13366)	62.9 (9556)	43.3 (10921)	16.8 (13002)	54.2 (4146)	46.3 (4146)	5.6 (4090)	2.7 (8285)	38.3 (7952)
HFT-1.5B-Think	87.8 (4379)	32.7 (13431)	75.0 (10181)	50.0 (10480)	16.8 (13628)	62.2 (5090)	52.9 (5090)	15.9 (1927)	9.3 (10616)	44.7 (8314)
LHRMs-1.5B	**87.8** (3722)	**35.3** (13491)	**75.0** (9065)	**50.4** (9490)	**17.2** (9342)	**61.1** (3103)	**54.0** (3103)	**16.9** (1250)	**10.4** (5289)	**45.3** (6428)
7B Size Models										
DeepSeek-R1-Distill-Qwen-7B	92.8 (3558)	55.5 (9488)	91.5 (6255)	58.1 (8635)	37.6 (11669)	74.3 (2824)	64.3 (2824)	19.1 (2209)	17.9 (5282)	56.8 (5860)
HFT-7B-Think	93.8 (3658)	56.7 (10778)	**95.0** (6456)	59.7 (8376)	38.4 (12046)	80.3 (3251)	68.9 (3251)	30.6 (1731)	23.3 (7442)	60.7 (6332)
LHRMs-7B	**93.8** (2616)	**66.7** (11031)	**95.0** (4976)	**61.2** (7540)	**38.8** (8432)	**81.5** (1906)	**69.6** (1906)	**35.0** (1086)	**26.0** (3416)	**63.1** (4768)

Fig.5 Accuracy and Output Length Comparison on Various Benchmarks

Thank you for your attention!