Towards Thinking-Optimal Scaling of Test-Time Compute for LLM Reasoning

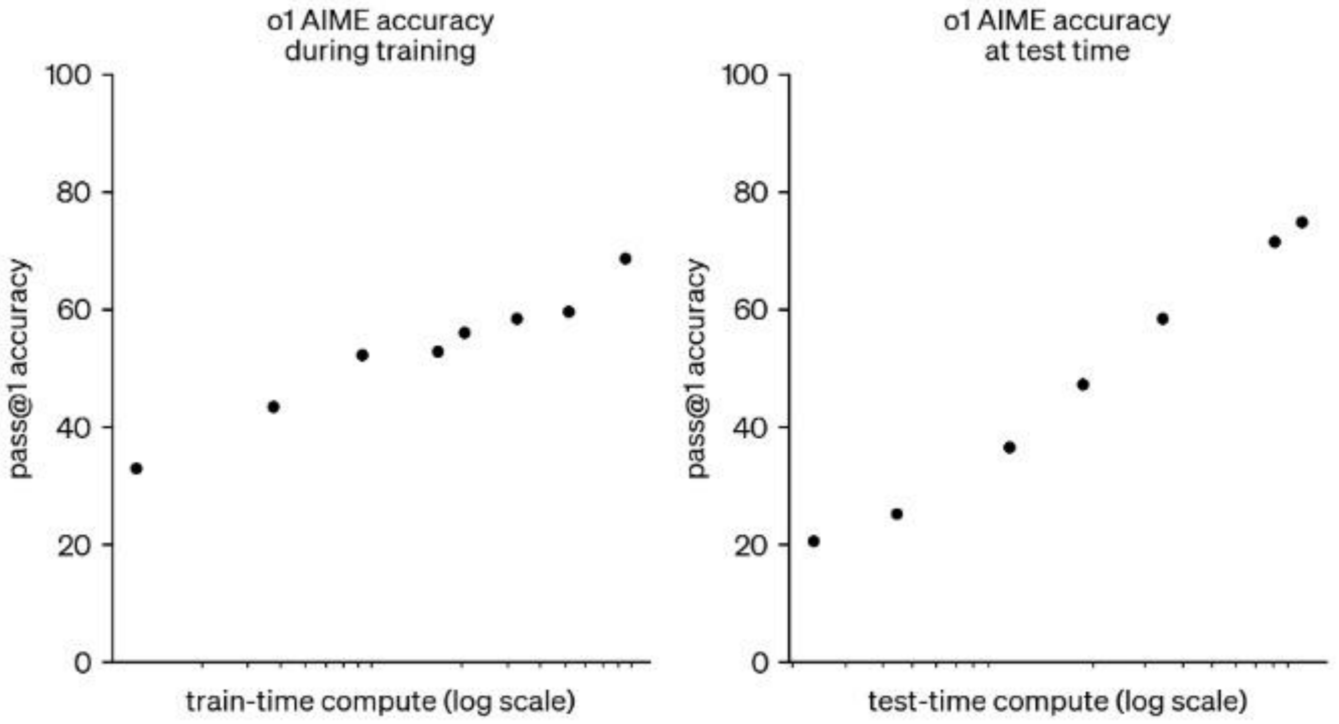
Wenkai Yang[1*], *Shuming Ma*[2], *Yankai Lin*[1#], *Furu Wei*[2]

[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Microsoft Research

*: Work done during an internship at Microsoft Research.
#: Corresponding Author

Research Problem

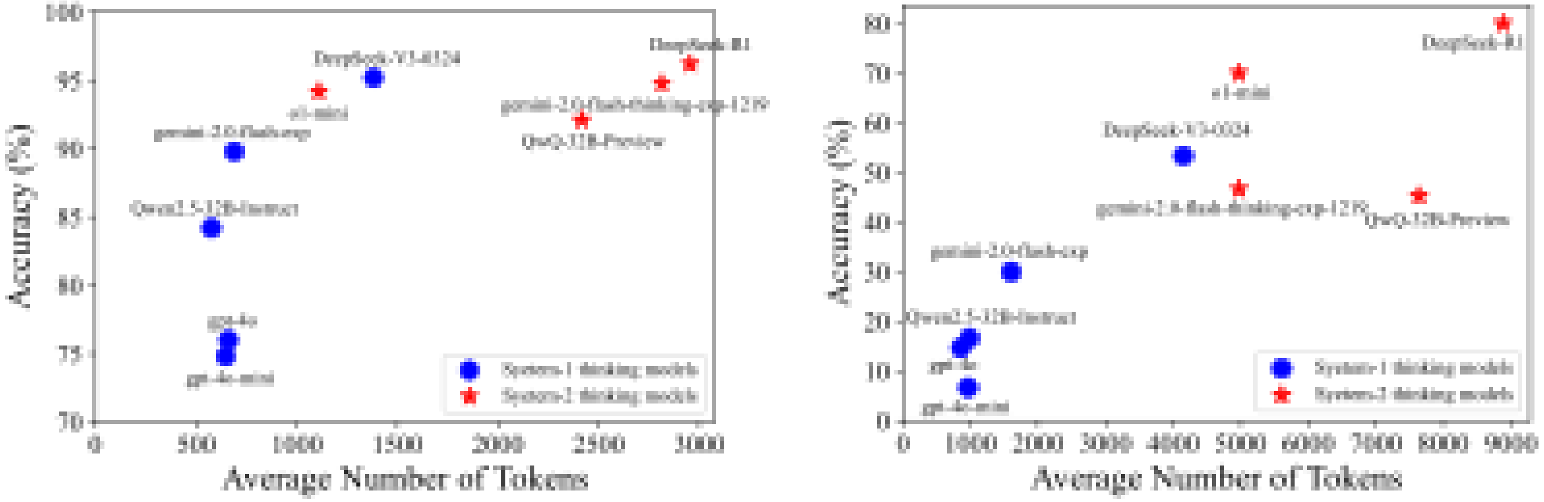
高瓴人工智能学院
Gaoling School of Artificial Intelligence



- Existing studies have shown that making a model spend more time thinking through longer Chain of Thoughts (CoTs) enables it to gain significant improvements in complex reasoning tasks.
- However, we are concerned about a potential issue hidden behind the current pursuit of test-time scaling: ***Would excessively scaling the CoT length actually bring adverse effects to a model's reasoning performance?***

[1] Learning to Reason with LLMs. OpenAI

Preliminary Analysis

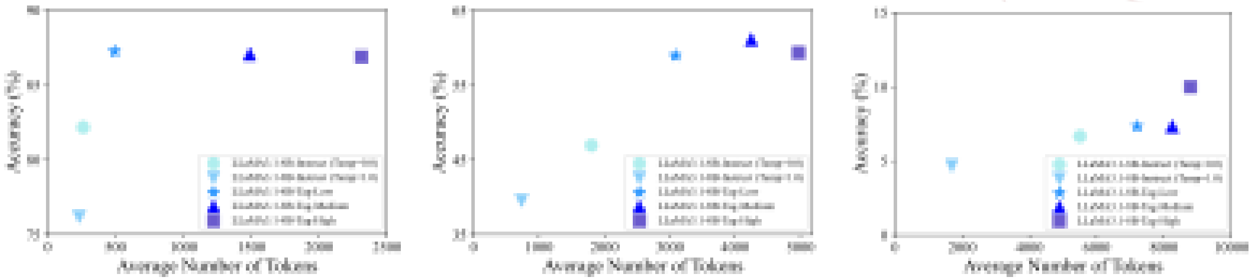


(a) Results on MATH500 (b) Results on AIME2024

Figure 1: The accuracy and the average number of tokens for each model on MATH500 and AIME2024. To ensure a fair comparison, we tokenized all model outputs using the Qwen2.5 tokenizer.

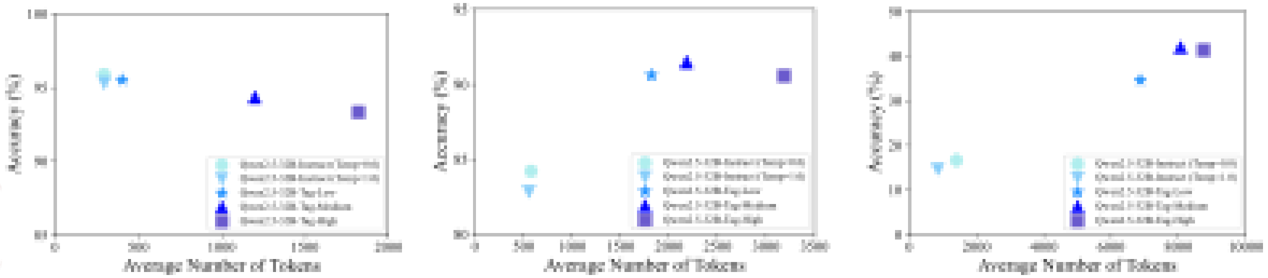
The preliminary analysis on several existing typical o1-like models along with their corresponding System-1 thinking models suggests, to some extent, that **excessively scaling to longer CoTs does not maximize test-time scaling effects.**

Deep Explorations



(a) Results on GSM8K (b) Results on MATH500 (c) Results on AIME2024

Figure 2: The accuracy and the average number of tokens of LLaMA3.1-8B-Instruct and LLaMA3.1-8B-Tag under different reasoning efforts ("Low", "Medium" and "High") on different benchmarks with varying levels of difficulty.



(a) Results on GSM8K (b) Results on MATH500 (c) Results on AIME2024

Figure 3: The accuracy and the average number of tokens of Qwen2.5-32B-Instruct and Qwen2.5-32B-Tag under different reasoning efforts ("Low", "Medium" and "High") on different benchmarks with varying levels of difficulty.

We fine-tune LLMs on a set of samples, where each problem is paired with three o1-like responses of different lengths, each assigned a distinct system prompt. Evaluation results show:

- **Scaling with longer CoTs can bring negative effects to the model's reasoning performance in certain domains**, especially on easy tasks.
- There exists an optimal reasoning efforts that varies across different tasks of varying difficulty levels.

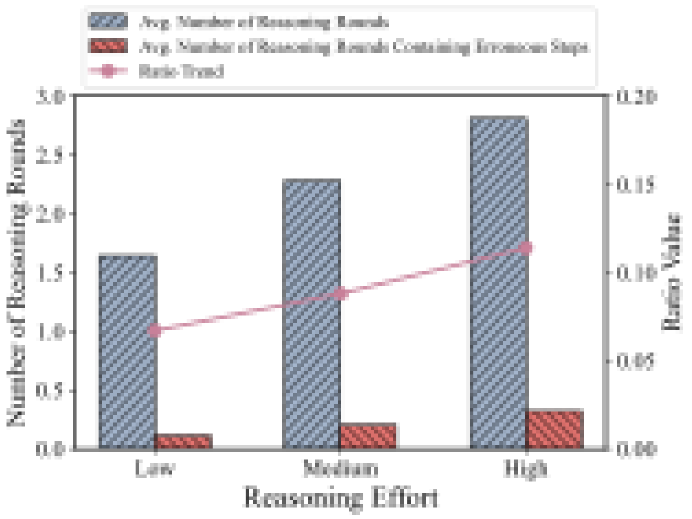
Deep Explorations

Table 10: The performance of Qwen2.5-7B-based models on MMLU-Pro and GPQA-Diamond

Model	MMLU-Pro		GPQA-Diamond	
	Accuracy	#Tokens	Accuracy	#Tokens
System-1 thinking models				
Qwen2.5-7B-Instruct (Temp. = 0.0)	52.46	401.38	34.85	592.73
Qwen2.5-7B-Instruct (Temp. = 1.0)	51.49	379.84	33.84	537.41
Tag models				
Qwen2.5-7B-Tag-General-Low	56.00	1674.60	31.82	2808.88
Qwen2.5-7B-Tag-General-Medium	55.92	2341.27	36.87	3931.13
Qwen2.5-7B-Tag-General-High	55.81	2632.05	32.83	4238.11

The above findings also hold true in the general reasoning domain as well.

Deep Explorations



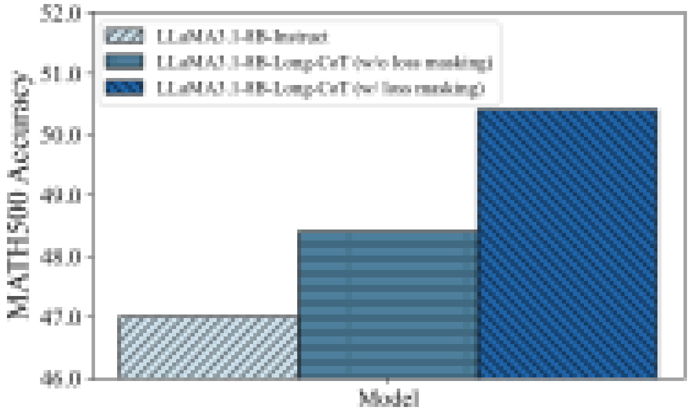


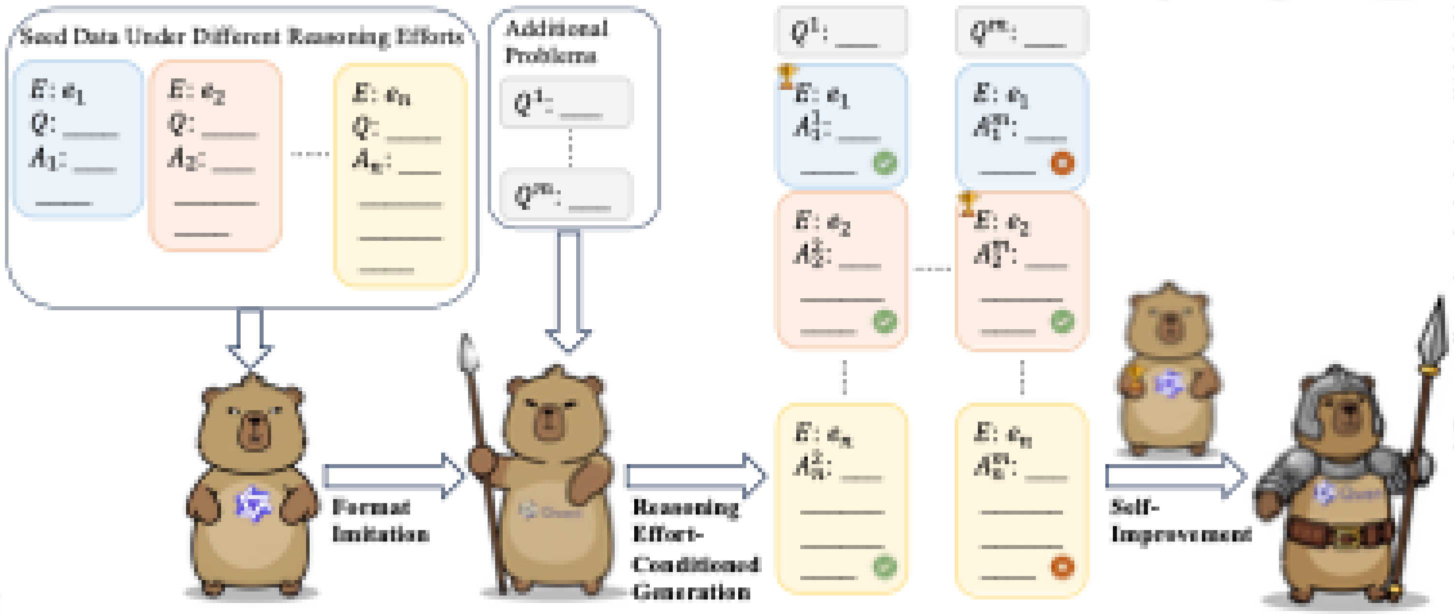
Figure 4: The statistics of responses under different reasoning efforts for training the tag models.

Figure 5: Empirical results of loss masking on erroneous steps. Evaluation temperature is 0.0.

- After analyzing the response properties, we find that **the number and proportion of erroneous reasoning rounds consistently increase as the reasoning effort grows**. Training the model on more wrong steps would bring adverse effects to the model's reasoning abilities, which can explain why scaling with higher reasoning effort leads to worse results.
- The additional results of applying loss masking to the tokens in the identified wrong steps of our custom-constructed long CoTs, as shown in Figure 5, help validate this claim.

Thinking-Optimal Test-Time Scaling

高瓴人工智能学院
Gaoling School of Artificial Intelligence



The illustration of our Thinking-Optimal Scaling method. Our method includes three stages:

- **Format Imitation** enables the base model to learn how to adopt different levels of reasoning effort to perform System-2 thinking, using a small set of seed data.
- **Reasoning Effort-Conditioned Generation** requires the model to apply System-2 thinking to a large set of problems under different reasoning efforts.
- **Self-Improvement** select the shortest correct response for each problem among all responses to fine-tune the base model to achieve thinking-optimal test-time scaling.

Main Results

Table 3: The results of our self-improved (Qwen2.5-32B-TOPS) and further iteratively self-improved models (Qwen2.5-32B-TOPS-Iter) compared to existing o1-like models using the same base model on GSM8K, MATH500, and AIME2024. In each setting, the underlined value represents the best result for System-1 thinking models, while the bold **value** indicates the best result for System-2 thinking models.

Model	GSM8K		MATH500		AIME2024	
	Accuracy	#Tokens	Accuracy	#Tokens	Accuracy	#Tokens
System-1 thinking models						
Qwen2.5-32B-Instruct (Temp. = 0.0)	95.91	295.01	84.20	576.89	16.67	1407.43
Qwen2.5-32B-Instruct (Temp. = 1.0)	95.30	296.98	82.84	555.65	14.67	855.62
System-2 thinking models						
QwQ-32B-Preview	95.23	761.01	**92.02**	2416.23	45.33	7636.63
STILL-2-32B	95.47	570.64	91.40	2005.28	45.33	6656.11
Sky-T1-32B-Preview	94.82	695.66	89.48	2022.07	35.33	5351.29
Qwen2.5-32B-Random	95.00	938.45	90.16	2670.19	39.33	7691.30
Qwen2.5-32B-TOPS (ours)	**95.82**	412.24	91.48	1883.29	43.33	7260.26
Qwen2.5-32B-TOPS-Iter-SFT (ours)	95.45	366.14	90.76	1701.11	44.00	6611.89
Qwen2.5-32B-TOPS-Iter-DPO (ours)	95.80	384.81	91.60	1731.72	**46.00**	6426.62

- The model trained under thinking-optimal samples (Qwen2.5-32B-TOPS) consistently performs better than the model trained under thinking-suboptimal samples (Qwen2.5-32B-Random), and outperforms other distillation-based models.
- The comparison of reasoning tokens used by different models across various domains reflects our model's ability to exhibit **adaptive reasoning depths**.
- Iterative self-improvement via DPO leads to continuous improvements in both efficiency and effectiveness.

Main Results

Table 4: The self-improvement results on LLaMA3.1-8B-Instruct. In each setting, the underlined value represents the best result for System-1 thinking models, while the bold **value** indicates the best result for System-2 thinking models.

Model	GSM8K		MATH500		AIME2024	
	Accuracy	#Tokens	Accuracy	#Tokens	Accuracy	#Tokens
System-1 thinking models						
LLaMA3.1-8B-Instruct (Temp. = 0.0)	82.18	262.23	47.00	1801.76	6.67	5506.30
LLaMA3.1-8B-Instruct (Temp. = 1.0)	76.21	233.08	39.60	733.56	4.67	1691.88
System-2 thinking models						
LLaMA3.1-8B-Random-SFT	87.94	1051.05	60.52	3627.23	4.67	8165.69
LLaMA3.1-8B-TOPS-SFT	**88.54**	571.10	**61.28**	3254.01	**8.00**	7392.59

The results on LLaMA3.1-8B-Instruct demonstrate the generalizability of our method on other architectures.

Thank you for listening!