Dynamic Early Exit in Reasoning Models

Chenxu Yang[1,2*], Qingyi Si[3*], Yongjie Duan[3], Zheliang Zhu[1,2], Chenyu Zhu[3], Qiaowei Li[3], Minghui Chen[1,2], Zheng Lin[1,2†], Weiping Wang[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China.
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China.
[3] Huawei Technologies Co., Ltd.

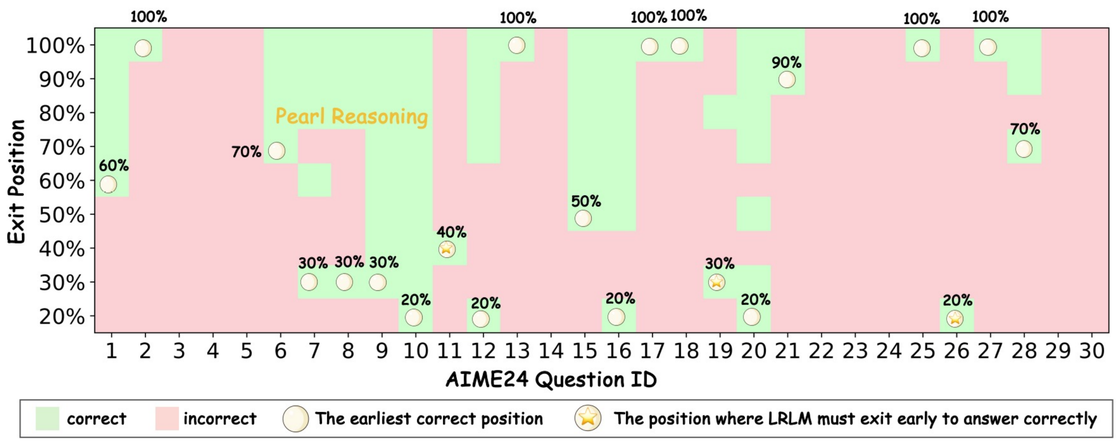
Email: yangchenxu@iie.ac.cn, linzheng@iie.ac.cn, siqingyi@huawei.com

Paper Link

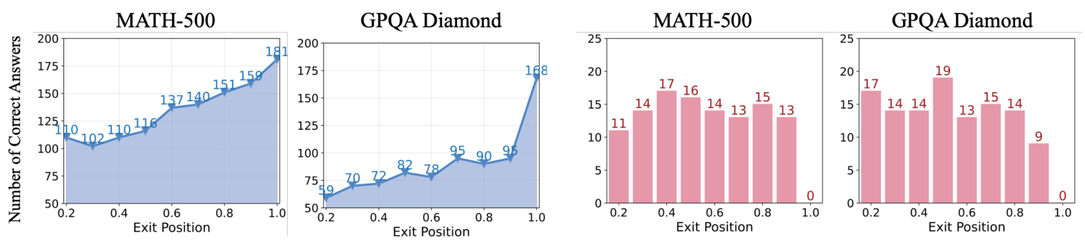
Introduction

- Recent advances in large reasoning language models (LRMs) rely on test-time scaling, which extends long chain-of-thought (CoT) generation to solve complex tasks. However, overthinking in long CoT not only slows down the efficiency of problem solving, but also risks accuracy loss due to the extremely detailed or redundant reasoning steps.
- we analyze the overthinking phenomenon in LRMs and investigate the impact of static early exits on model performance. We define ”pearl reasoning” as the critical juncture where reasoning information becomes precisely sufficient for accurate problem-solving and verify the existence of such pearl reasoning.
- We propose a simple yet effective method that allows LLMs to self-truncate CoT sequences by early exit during generation. Instead of relying on fixed heuristics, the proposed method monitors model behavior at potential reasoning transition points and dynamically terminates the next reasoning chain's generation when the model exhibits high confidence in a trial answer.

Analysis of Pearl Reasoning



Correctness statistics for early exits at various reasoning steps.

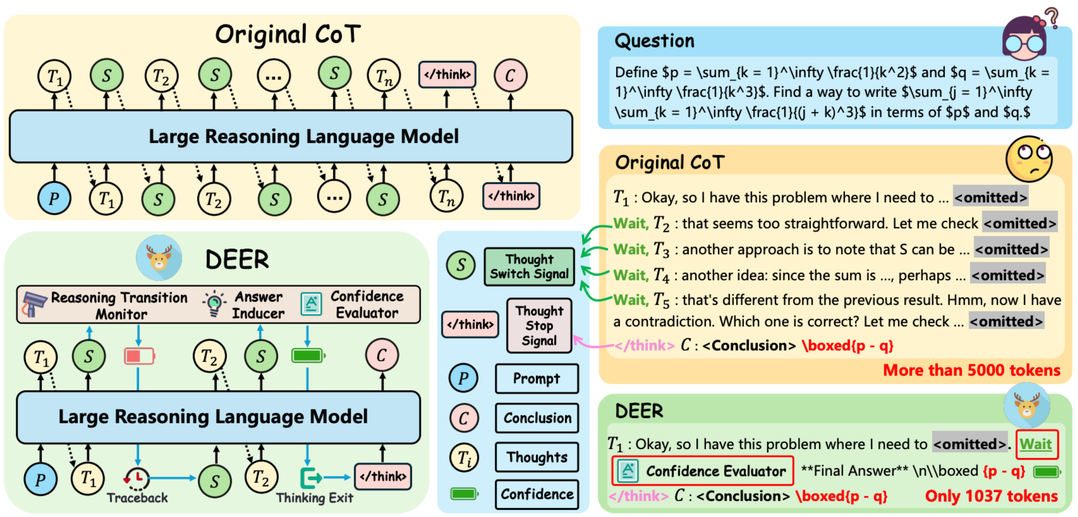


(a) The number of **originally correct** samples that remain correct with early exiting on different positions.

(b) The number of **originally incorrect** samples that become correct with early exiting on different positions.

LRMs possess the potential to achieve simultaneous improvements in both computational efficiency and prediction accuracy through strategic early termination.

Method



An overview of the Dynamic Early Exit in Reasoning (DEER) method.

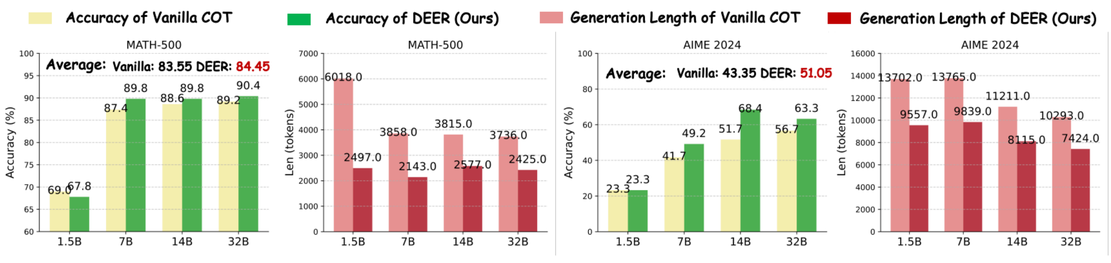
The core idea behind DEER is that a model's confidence in its trial answer dynamically indicates whether the thinking information required for LRMs to generate the final answer is sufficient.
DEER involves three designs to determine whether to exit early: reasoning transition monitor, answer inducer, and confidence evaluator.

Experiments

Main Experiments

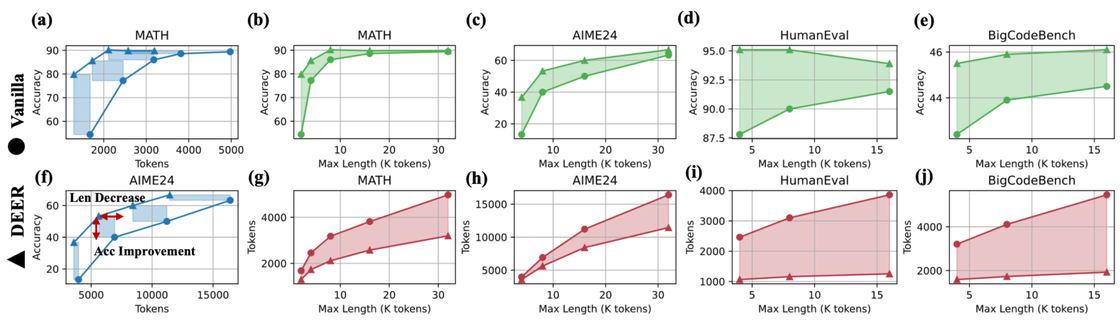
Method	Math: GSM8K Acc↑	GSM8K Tok↓	GSM8K CR↓	MATH-500 Acc↑	MATH-500 Tok↓	MATH-500 CR↓	AMC23 Acc↑	AMC23 Tok↓	AMC23 CR↓	AIME24 Acc↑	AIME24 Tok↓	AIME24 CR↓	Science: GPQA-D Acc↑	GPQA-D Tok↓	GPQA-D CR↓	Overall Acc↑	Overall CR↓
DeepSeek-R1-Distill-Qwen-7B																	
Vanilla	89.6	1,484	100%	87.4	3,858	100%	78.8	6,792	100%	41.7	13,765	100%	23.7	10,247	100%	64.2	100%
TCC	88.0	892	60.1%	89.2	3,864	100.2%	82.5	6,491	95.6%	48.4	10,603	77.0%	27.3	8,442	82.4%	67.1	83.0%
CoD	84.7	298	20.1%	83.2	1,987	51.5%	77.5	4,440	65.4%	40.0	10,519	76.4%	37.9	6,431	62.8%	64.7	55.3%
NoThinking	87.1	284	19.1%	80.6	834	21.6%	65.0	1,911	28.1%	26.7	4,427	32.2%	29.8	724	7.1%	57.8	**21.6%**
Dynasor-CoT	89.6	1,285	86.6%	89.0	2,971	77.0%	85.0	5,980	88.0%	46.7	12,695	92.2%	30.5	7,639	74.5%	68.2	83.7%
SEAL	88.4	811	54.6%	89.4	2,661	69.0%	–	–	–	–	–	–	–	–	–	–	–
DEER	90.6	917	61.8%	89.8	2,143	55.5%	85.0	4,451	65.5%	49.2	9,839	71.5%	31.3	5,469	53.4%	69.2	61.5%
DEER-PRo	91.0	989	66.7%	90.2	2,391	62.0%	87.5	4,877	71.8%	49.2	10,046	73.0%	30.6	5,682	55.5%	**69.7**	65.8%
Qwen3-14B																	
Vanilla	95.1	2,047	100%	93.8	4,508	100%	95.0	7,139	100%	70.0	10,859	100%	60.1	7,339	100%	82.8	100%
TCC	95.7	1,241	60.6%	94.6	4,484	99.5%	95.0	7,261	101.7%	70.8	11,573	106.6%	60.1	7,138	97.3%	83.3	93.1%
CoD	85.7	648	31.7%	75.2	2,359	52.3%	72.5	4,122	57.7%	60.0	10,768	99.2%	51.0	1,177	16.0%	68.9	51.4%
NoThinking	94.8	286	14.0%	85.0	1,228	27.2%	77.5	2,133	29.9%	26.7	7,337	67.6%	50.5	2,320	31.6%	66.9	**34.1%**
Dynasor-CoT	95.6	1,483	72.4%	93.8	4,063	90.1%	95.6	6,582	92.2%	73.3	10,369	95.5%	59.6	5,968	81.3%	83.6	86.3%
DEER	95.3	840	41.0%	94.0	3,074	68.2%	95.0	4,763	66.7%	76.7	7,619	70.2%	57.6	2,898	39.5%	83.7	57.1%
DEER-PRo	95.3	926	45.2%	94.4	3,260	72.3%	95.6	4,905	68.7%	75.0	8,135	74.9%	61.2	4,062	55.4%	**84.3**	63.3%
QwQ-32B																	
Vanilla	96.7	1,427	100%	93.8	4,508	100%	92.5	6,792	100%	66.7	10,821	100%	63.1	7,320	100%	82.6	100%
TCC	95.8	1,348	94.5%	94.4	4,315	95.7%	90.0	6,818	100.4%	60.0	11,263	104.1%	61.6	7,593	103.7%	80.4	99.7%
CoD	96.0	627	43.9%	94.0	3,630	80.5%	92.5	5,943	87.5%	60.0	10,731	99.2%	62.6	6,039	82.5%	81.0	**78.7%**
NoThinking	96.2	1,113	78.0%	94.8	3,930	87.2%	87.5	6,908	101.7%	66.7	10,859	100.4%	63.6	7,668	104.8%	81.8	94.4%
Dynasor-CoT	95.2	1,095	76.7%	94.2	4,176	92.6%	93.8	6,544	96.3%	63.3	11,156	103.1%	64.1	7,024	96.0%	82.1	93.0%
DEER	96.3	977	68.5%	94.6	3,316	73.6%	95.0	5,782	85.1%	70.0	10,097	93.3%	64.1	6,163	84.2%	84.0	80.9%
DEER-PRo	96.2	1032	72.3%	94.8	3,650	80.9%	95.0	5,811	85.6%	70.0	10,264	94.9%	64.7	6,201	84.7%	**84.1**	83.7%

Experimental results across various types of reasoning models.

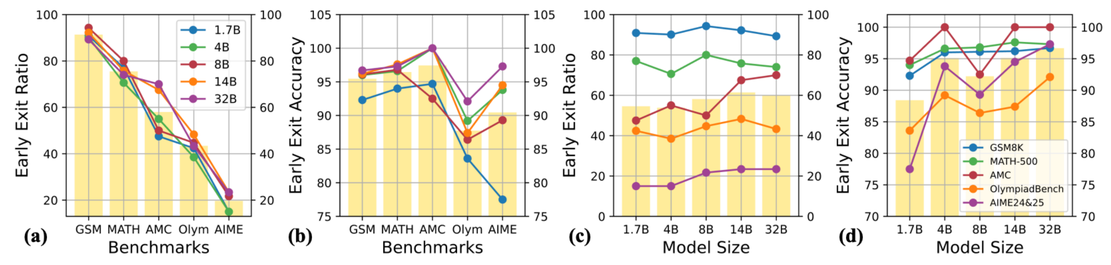


Experimental results across DeepSeek-R1-DistillQwen-Series models of varying sizes.

Analysis



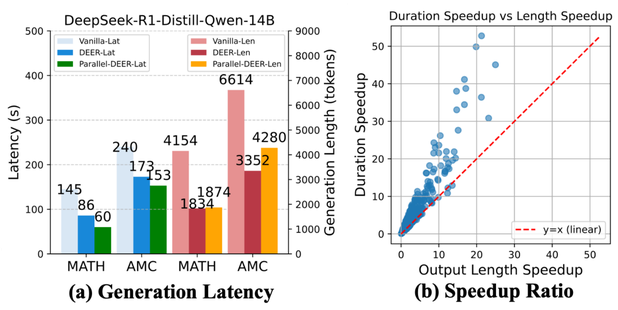
Performance across four datasets under different token budget settings.



Early-exit rates and accuracy of early-exited samples of DEER.

Efficiency Improvement

Branch-Parallel DEER achieves further speed improvements by efficiently reducing the latency of trial answer inducing and confidence evaluation. The ratio between latency speedup and length savings exhibits a superlinear trend, reinforcing the significance of DEER in enhancing inference speed.



Conclusion

This paper verifies the rationale behind the early exit motivation in CoT generation, and accordingly proposes a training-free dynamic early exit algorithm, which makes the reasoning model withdraw from subsequent thinking when the thinking amount is just enough. Our method comprehensively evaluated across reasoning models of varying model sizes and demonstrates superior performance with fewer tokens.