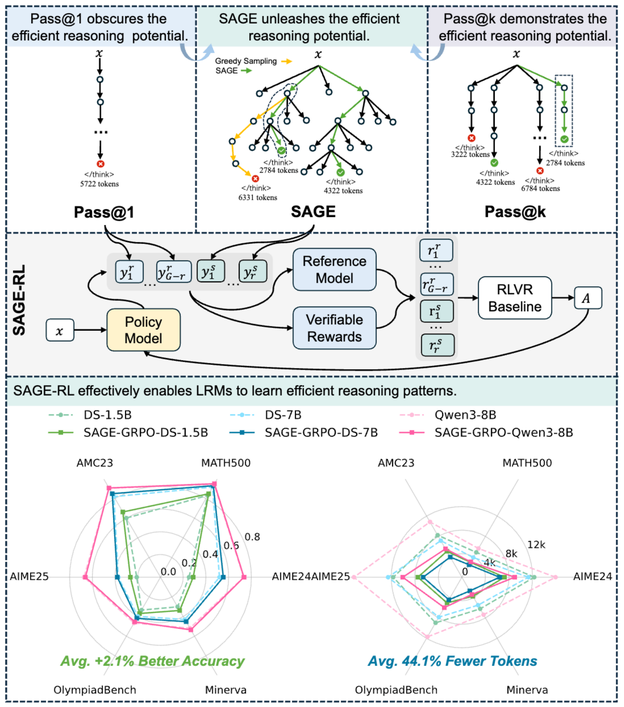
Does Your Reasoning Model Implicitly Know When to Stop Thinking?

Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuanda Wang, Zhixia Zhang, Hongyan Xie,
Songshi Liang, Zehao Chen, Xuefeng Xiao, Fuzhen Zhuang, Jianxin Li, Deqing Wang, Yikun Ban

ICML International Conference On Machine Learning

MOTIVATION

- **Pass@1:** Existing sampling strategies fail to enable timely termination of thinking.
- **Pass@k:** Scaling CoT length does not lead to correct answers.



The surprising performance of relatively shorter responses in pass@k reveals the inherent potential of the model for efficient reasoning. The pervasive redundancy of reasoning steps in pass@1 indicates that current sampling paradigms obscure this potential.

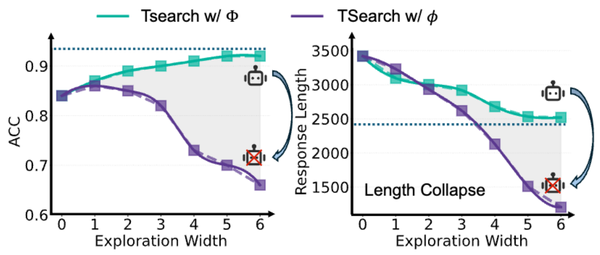
METHODOLOGY

TSearch uses the average cumulative log-probability Φ as a confidence measure over reasoning paths, performs token-level expansion to retain the top-m high-confidence candidate chains, and terminates once </think> appears with sufficiently high confidence, thereby revealing the existence of concise yet effective reasoning paths within the model.

SAGE further reformulates TSearch from token-level search into step-level reasoning-chain exploration: it samples complete reasoning steps at each expansion stage and directly leverages the model's confident termination signal over </think> to identify shorter and more accurate CoTs.

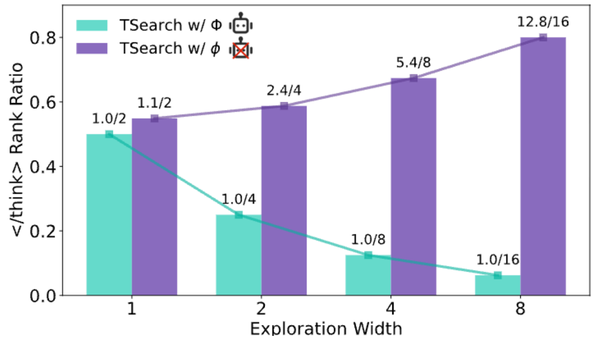
SAGE-RL incorporates SAGE as a hybrid sampler in the RLVR rollout process, where part of each response group is generated by SAGE as high-quality concise reasoning samples while the remainder is produced by standard random sampling, enabling the model to internalize efficient reasoning patterns under standard pass@1 inference.

OBSERVATIONS

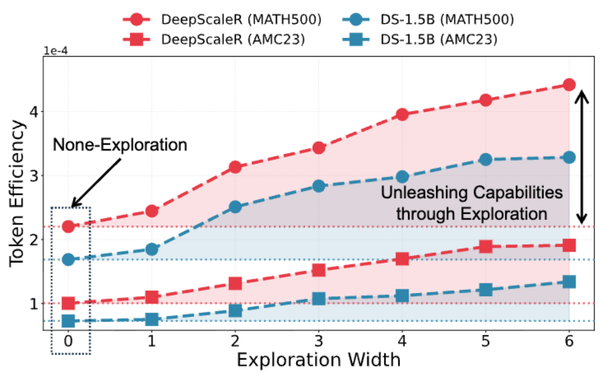


High-Confidence Paths Lead to Efficient Reasoning

Method	TR	ACC	T-LEN	LEN
Random	-	0.84	3126	3419
TSearch (4,1) w/ ϕ	1.00	0.79	1712	2129
TSearch (4,1) w/ ϕ	0.75	0.82	2022	2333
TSearch (4,1) w/ ϕ	0.50	0.89	2176	2609
TSearch (4,1) w/ Φ	1.00	0.92	2213	2609
TSearch (4,1) w/ Φ	0.75	0.92	2221	2621
TSearch (4,1) w/ Φ	0.50	0.91	2212	2632



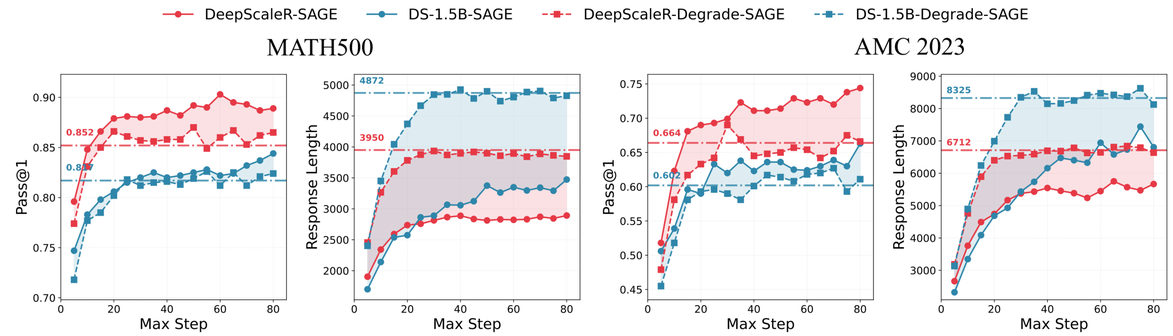
High-Confidence Paths Lead to Confident Ends



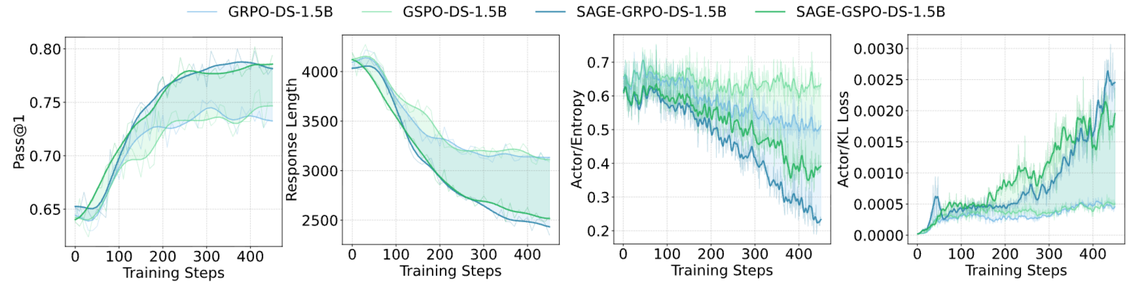
Scaling Exploration Drives Capability Convergence

EXPERIMENTS

Performance of SAGE



Training Dynamics of SAGE-RL



Performance on Challenge Math Benchmarks

Method	MATH-500			AIME 2024			AIME 2025			OlympiadBench		
	Pass@1↑(%)	LEN↓	TE↑($\times10^{-3}$)	Pass@1↑(%)	LEN↓	TE↑($\times10^{-3}$)	Pass@1↑(%)	LEN↓	TE↑($\times10^{-3}$)	Pass@1↑(%)	LEN↓	TE↑($\times10^{-3}$)
DS-1.5B	83.2	4882	17.0	25.1	12300	2.04	20.9	11669	1.79	33.4	8954	3.73
+ LC-R1	80.4 (↓2.8)	2973 (↓1909)	27.0 (↑58.8%)	23.3 (↓1.8)	7098 (↓5202)	3.28 (↑60.8%)	20.9 (↑0.0)	6942 (↓4727)	3.01 (↑68.2%)	32.0 (↓1.4)	4632 (↓4322)	6.91 (↑85.3%)
+ ThinkPrune-2k	81.7 (↓1.5)	2826 (↓2056)	28.9 (↑70.0%)	23.7 (↓1.4)	7085 (↓5215)	3.35 (↑64.2%)	19.7 (↓1.2)	**6918** (↓4751)	2.85 (↑59.2%)	32.9 (↓0.5)	4752 (↓4202)	6.92 (↑85.5%)
+ AdaptThink	80.4 (↓2.8)	**2563** (↓2319)	**31.4** (↑84.1%)	25.7 (↑0.6)	8055 (↓4245)	3.19 (↑56.4%)	21.8 (↑0.9)	8155 (↓3514)	2.67 (↑49.2%)	32.6 (↓0.8)	**4563** (↓4391)	7.14 (↑91.4%)
+ Efficient Reasoning	82.0 (↓1.2)	2821 (↓2061)	29.1 (↑70.6%)	26.2 (↑1.1)	9189 (↓3111)	2.85 (↑39.7%)	22.9 (↑2.0)	8590 (↓3079)	2.67 (↑49.2%)	33.8 (↑0.4)	5755 (↓3199)	5.87 (↑57.4%)
+ GRPO	83.6 (↑0.4)	3907 (↓975)	21.4 (↑25.6%)	28.3 (↑3.2)	8767 (↓3533)	3.23 (↑58.3%)	24.1 (↑3.2)	8263 (↓3406)	2.92 (↑63.1%)	34.2 (↑0.8)	6323 (↓2631)	5.41 (↑45.0%)
+ **SAGE-GRPO**	84.8 (↑1.6)	2915 (↓1967)	29.1 (↑70.7%)	**28.8** (↑3.7)	7243 (↓5057)	3.98 (↑95.1%)	26.5 (↑5.6)	7479 (↓4190)	3.54 (↑97.8%)	36.9 (↑3.5)	5050 (↓3904)	**7.31** (↑96.0%)
+ GSPO	83.4 (↑0.2)	3898 (↓984)	25.3 (↑21.4%)	28.3 (↑3.2)	8604 (↓3696)	3.29 (↑61.3%)	25.1 (↑4.2)	8227 (↓3442)	3.05 (↑70.4%)	34.6 (↑1.2)	6410 (↓2544)	5.40 (↑44.8%)
+ **SAGE-GSPO**	**85.2** (↑2.0)	2921 (↓1961)	29.2 (↑71.6%)	28.5 (↑3.4)	**6889** (↓5411)	**4.14** (↑102.9%)	**27.1** (↑6.2)	7167 (↓4502)	**3.78** (↑111.1%)	**37.3** (↑3.9)	5172 (↓3782)	7.21 (↑93.3%)
DeepScaleR	86.0	3805	22.6	31.4	9370	3.35	25.4	9310	2.73	35.9	5972	6.01
+ ThinkPrune-2k	82.5 (↓3.5)	**2946** (↓859)	28.0 (↑23.9%)	33.5 (↑2.1)	8108 (↓1262)	4.13 (↑23.3%)	26.0 (↑0.6)	**7486** (↓1824)	3.47 (↑27.1%)	35.1 (↓0.8)	**4723** (↓1249)	7.43 (↑23.6%)
+ GRPO	87.6 (↑1.6)	3482 (↓323)	25.2 (↑11.3%)	35.6 (↑4.2)	8592 (↓778)	4.14 (↑23.6%)	**27.4** (↑2.0)	8185 (↓1125)	3.35 (↑22.7%)	36.2 (↑0.3)	5443 (↓529)	6.65 (↑10.6%)
+ **SAGE-GRPO**	**88.8** (↑2.8)	3117 (↓688)	**28.4** (↑25.7%)	**36.1** (↑4.7)	**8094** (↓1276)	**4.46** (↑33.1%)	27.2 (↑1.8)	7704 (↓1606)	**3.53** (↑29.3%)	**36.5** (↑0.6)	4890 (↓1082)	**7.46** (↑24.1%)
DS-7B	91.6	3871	23.7	51.9	11305	4.59	37.1	12540	2.96	39.8	7839	5.08
+ LC-R1	87.3 (↓4.3)	**2076** (↓1795)	42.1 (↑77.7%)	51.7 (↓0.2)	6820 (↓4485)	7.58 (↑65.1%)	35.7 (↓1.4)	7458 (↓5082)	4.79 (↑61.8%)	41.4 (↑1.6)	**4193** (↓3646)	**9.87** (↑94.3%)
+ AdaptThink	88.9 (↓2.7)	2199 (↓1672)	40.4 (↑70.9%)	52.1 (↑0.2)	6679 (↓4626)	7.80 (↑69.9%)	35.0 (↓2.1)	7807 (↓4733)	4.48 (↑72.3%)	38.9 (↓0.9)	4915 (↓2924)	7.91 (↑55.7%)
+ Efficient Reasoning	89.8 (↓1.8)	2408 (↓1463)	37.3 (↑57.6%)	51.9 (↑0.0)	6667 (↓4638)	7.78 (↑69.5%)	36.2 (↓0.9)	7501 (↓5039)	4.82 (↑62.8%)	40.1 (↑0.3)	4599 (↓3240)	8.72 (↑71.7%)
+ GRPO-LEAD	89.5 (↓2.1)	2752 (↓1119)	32.5 (↑37.1%)	53.1 (↑1.2)	7023 (↓4282)	7.56 (↑64.7%)	36.1 (↓1.0)	7842 (↓4698)	4.60 (↑55.4%)	40.6 (↑0.8)	4972 (↓2867)	8.17 (↑60.8%)
+ GRPO	92.0 (↑0.4)	3219 (↓652)	28.5 (↑20.2%)	52.5 (↑0.6)	8424 (↓2881)	6.23 (↑35.7%)	**38.4** (↑1.3)	10123 (↓2417)	3.79 (↑28.0%)	41.2 (↑1.4)	5498 (↓2341)	7.50 (↑47.6%)
+ **SAGE-GRPO**	**93.0** (↑1.4)	2141 (↓1730)	**43.4** (↑83.1%)	**55.3** (↑3.4)	**6422** (↓4883)	**8.61** (↑87.6%)	38.0 (↑0.9)	**6583** (↓5957)	**5.77** (↑94.9%)	**41.8** (↑2.0)	4435 (↓3404)	9.42 (↑85.4%)
Qwen3-8B	94.4	5640	16.7	73.2	15920	4.60	**67.3**	18342	3.67	46.6	11707	4.00
+ GRPO	93.6 (↓0.8)	4470 (↓1170)	20.9 (↑25.1%)	72.8 (↓0.4)	10573 (↓5347)	6.89 (↑49.8%)	66.6 (↓0.7)	13981 (↓4361)	4.76 (↑29.7%)	45.1 (↓1.5)	7512 (↓4195)	6.00 (↑50.0%)
+ **SAGE-GRPO**	**95.0** (↑0.6)	3015 (↓2625)	31.5 (↑88.2%)	73.5 (↑0.3)	8975 (↓6945)	8.19 (↑78.0%)	66.6 (↓0.7)	10052 (↓8290)	6.58 (↑79.3%)	45.4 (↓1.2)	5972 (↓5735)	7.60 (↑90.0%)
+ GSPO	94.6 (↑0.2)	4342 (↓1298)	22.2 (↑32.9%)	73.0 (↓0.2)	10544 (↓5376)	6.92 (↑50.4%)	66.2 (↓1.1)	14082 (↓4260)	4.70 (↑30.2%)	46.6 (↑0.0)	7964 (↓3743)	5.85 (↑46.2%)
+ **SAGE-GSPO**	94.4 (↑0.0)	**2753** (↓2887)	**34.3** (↑105.3%)	**73.7** (↑0.5)	**8547** (↓7373)	**8.62** (↑87.4%)	66.0 (↓1.3)	**9183** (↓9159)	**7.19** (↑95.9%)	**46.7** (↑0.1)	**5436** (↓6271)	**8.59** (↑114.7%)

Statistics of RFCS on MATH500 across LRMs