Synergizing Parallel and Sequential Reasoning via Semantic Entropy-Guided Adaptive Termination

Anonymous ACL submission

Abstract

Test-time scaling has emerged as a critical driver for advancing Large Language Model (LLM) reasoning, yet current approaches remain bifurcated between sequential scaling and parallel scaling. Sequential methods often struggle with fixed token budgets, leading to premature halting or verbosity, while parallel methods typically lack inter-path coordination. To bridge this gap, we propose SEAT (Semantic Entropy-Guided Adaptive Termination), a training-free framework that synergizes the benefits of both paradigms. Specifically, SEAT adopts a hybrid architecture that simultaneously explores multiple reasoning paths while sequentially feeding results from the previous round into the next to refine the generation process. Our approach is grounded in the observation that Semantic Entropy (SE) strongly correlates negatively with model accuracy, serving as a reliable proxy for reasoning quality. SEAT leverages this signal to dynamically control the reasoning process, employing a novel threshold-free termination mechanism inspired by the "Secretary Problem" in Optimal Stopping Theory to eliminate pre-sampling overhead. Extensive evaluations across five challenging reasoning benchmarks demonstrate that SEAT significantly boosts performance. Furthermore, our adaptive approach effectively prevents semantic entropy collapse found in smaller 7B models, ensuring robust multi-round reasoning.

1 Introduction

Recent advances in large language models (LLMs), exemplified by models such as o1 (OpenAI, 2024), DeepSeek-R1 (DeepSeek-AI et al., 2025), and QwQ (Team, 2025), have significantly accelerated progress toward artificial general intelligence. A key driver is test-time scaling (Snell et al., 2025), which enhances performance by allocating more computational budget for in-depth reasoning. Current approaches generally fall into two categories:

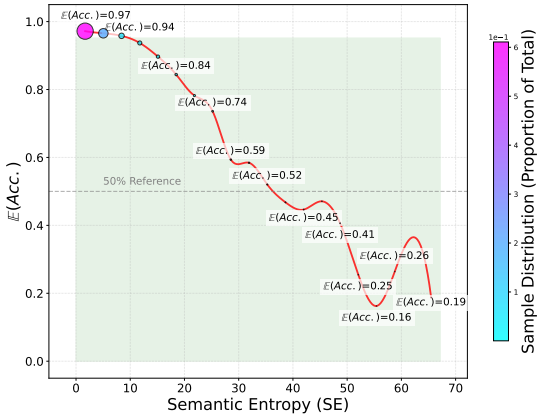


Figure 1: Strong negative correlation between semantic entropy and R1-Distill-Qwen-7B performance on MATH-500 benchmark under $N = 8$ parallel inferences per sample. The $\mathbb{E}(Acc.)$ denotes the expected accuracy and is calculated as the proportion of correct responses per N-inference set.

sequential scaling and parallel scaling. Sequential scaling typically explores iterative refinement via multi-round prompting (Muennighoff et al., 2025) or by steering generation with special tokens (e.g., "wait") (Muennighoff et al., 2025). However, without external supervision, they typically rely on fixed computational budgets, often leading to either unnecessary verbosity or premature halting (Chen et al., 2024; Wang et al., 2025c).In contrast, parallel scaling methods, such as best-of-N sampling (Cobbe et al., 2021; Kang et al., 2025) and majority voting (Wang et al., 2022), promote diverse exploration through independent sampling but typically lack inter-path coordination. In this work, we investigate the following research question: **How can we design a flexible framework to synergize the benefits of sequential and parallel scaling?**

The primary challenge in achieving this synergy lies in adaptively controlling sequential scaling. We address this by leveraging uncertainty from parallel scaling as a performance proxy. Intuitively,

high semantic diversity implies the model is struggling (Chen et al., 2023; Liang et al., 2024; Zeng et al., 2025), suggesting that extending the reasoning process until this uncertainty diminishes could improve quality. To operationalize this, we adopt semantic entropy (SE) (Malinin and Gales, 2021; Chen et al., 2025) as a metric to quantify reasoning quality. As shown in Fig. 1, experiments on MATH-500 reveal a strong negative correlation between SE and accuracy, with accuracy dropping from 97% to 19% as SE increases. This demonstrates that **semantic entropy is an effective signal for controlling reasoning, enabling a synergy between parallel and sequential strategies**.

Based on the above insights, we propose SEAT[1], a universal, training-free hybrid reasoning framework. As shown in Fig. 2, SEAT adopts a hybrid architecture that simultaneously explores multiple reasoning paths, while sequentially feeding the results from the previous round into the next to refine the generation process. For parallel scaling, it dynamically adjusts the degree of parallelization to expand exploration and prevent the model from getting stuck on local optima. Simultaneously, for sequential refinement, it leverages semantic entropy to trigger early stopping, thereby reducing wasted computation. To determine when to terminate this process, we first establish a threshold-based baseline. To find an appropriate stopping threshold for a given N (N is the pre-defined number of parallel responses per round), Statistical analysis reveals a similar $80/20$ pattern (Wang et al., 2025a): 80% of correct answers had been selected from the lowest 20% of the SE distribution. Therefore, we define the threshold using the SE value at the 20th percentile. However, to eliminate pre-sampling overhead and enhance robustness, we propose a threshold-free, adaptive variant. Inspired by the 'Secretary Problem' in Optimal Stopping Theory (Ferguson, 1989; Shiryaev, 1980), this mechanism establishes a dynamic baseline from the model's initial reasoning steps. The process terminates immediately when subsequent semantic entropy falls below this baseline, ensuring efficiency without static thresholds.

We evaluate SEAT across model sizes from 7B to 32B on five demanding benchmarks: AIME-2024, AIME-2025, MATH-500, MINERVA, and GPQA. Experimental results show that our adaptive method substantially improves performance. Notably, even with as few as $N = 2$ parallel responses, the 32B and 7B models achieve accuracy gains of 14.1% and 24.5% on AIME-2025. Moreover, our framework demonstrates excellent extensibility, achieving even higher accuracy when combined with strategies like max probability and majority voting. Surprisingly, we find that our adaptive approach effectively prevents semantic entropy collapse in smaller 7B models. This notorious phenomenon, characterized by a sudden drop in entropy after several parallelization steps, typically traps the model in a loop of over-confidently repeating incorrect answers. By adaptively terminating the reasoning process before this collapse occurs, our method enables 7B models to sustain performance across multi-round parallelization.

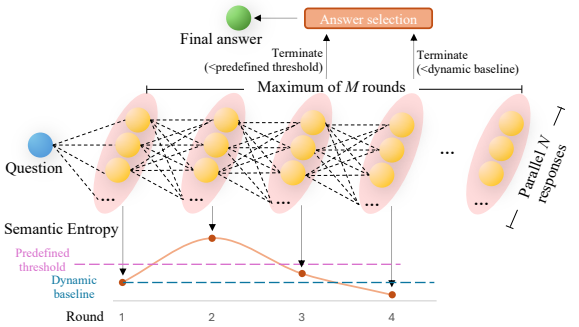


Figure 2: The overview of our proposed SEAT.

2 Methodology

2.1 Method Overview

Given a question, this paper employs multi-round parallel reasoning to generate a collection of candidate responses, and then selects the final answer from this set. Three primary components are involved in the above procedure: (1) the design of the multi-round parallel reasoning framework, (2) calculation of the SE metric during inference with termination mechanism, and (3) selection strategy for the final answer among candidate responses. These components will be described step by step.

Multi-round Parallel (MRP) Inference Framework. As illustrated in Fig. 2, the proposed SEAT framework establishes an $N \times M$ reasoning structure for multi-round parallel reasoning, where N represents the parallel dimension (*i.e.*, number of reasoning paths per round) and M is the sequential dimension (*i.e.*, number of reasoning rounds). Notably, $N \times M$ serves as the pre-defined maximum reasoning budget allocated to the model. Due

[1]SEAT stands for a **S**emantic **E**ntropy-Guided **A**daptive **T**ermination Framework

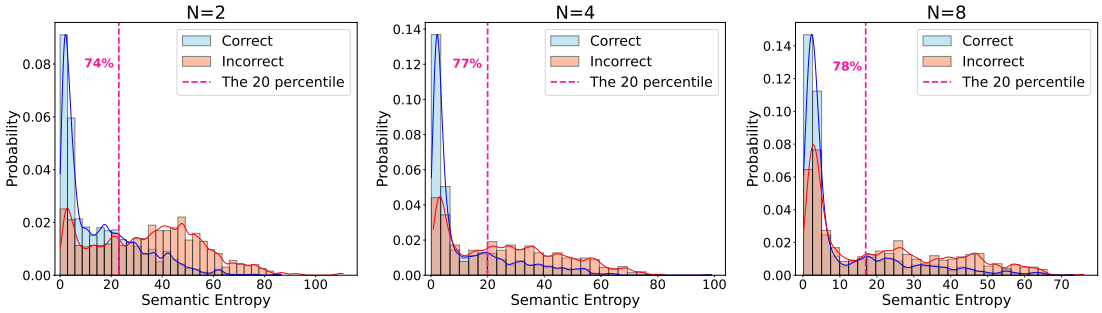


Figure 3: Semantic entropy distribution of correct and incorrect answers. The lowest 20% threshold is marked by red line and the proportion of correct answers within this region is labeled numerically in red.

to our proposed adaptive termination mechanism during inference, the actual computational budget typically falls below this pre-defined maximum. In contrast to prior test-time scaling approaches, SEAT synergistically integrates sequential refinement and parallel exploration. Specifically, each sequential round can access all N reasoning outputs from the prior round, enabling the model to refine its reasoning by leveraging diverse responses for error correction. Furthermore, within each round, all parallel reasoning paths operate independently, which aims to increase the diversity and encourage exploration. Formally, the j-th reasoning path in the i-th round is defined as:

$$MRP_j^i(P_i) \rightarrow \{\text{Thinking}_j^i, \text{Answer}_j^i\}, \quad (1)$$

where $i \in \{1, \cdots, M\}$ and $j \in \{1, \cdots, N\}$. Here, P_i denotes the input prompt for the i-th round, which is initialized as the user prompt (i.e., $P_1 = \langle\text{user prompt}\rangle$). For every round from the second onward, we extract the answer segments of all N candidate responses produced by the parallel reasoning paths in the preceding round, and incorporate these N answers into the prompt for the current round. The prompt template can be found in the Appendix. By constructing the prompt for the i-th round using N parallel answers generated in the previous round, this approach encourages the model to review and refine its previous outputs for improved responses.

Semantic Entropy Calculation and Termination Mechanism. Given a question q and N answers $\mathcal{A} \triangleq \{a_1, a_2, \ldots, a_N\}$ extracted from the previous round's N responses, we compute the semantic entropy to quantify the model's uncertainty about q in light of these answers. Since the semantic entropy computation process is applied across different reasoning rounds, we omit the round-indicating superscripts for simplicity. The semantic entropy is defined as follows:

$$\text{SE}(q) = -\sum_c \left[\left(\sum_{r \in c} p\left(r \mid \{q; \mathcal{A}\}\right) \right) \log \left(\sum_{r \in c} p\left(r \mid \{q; \mathcal{A}\}\right) \right) \right], \quad (2)$$

where c denotes a possible semantic meaning class and r denotes a possible response. It's intractable to enumerate every possible c since LLMs can generate an unlimited number of diverse responses for a given question, potentially spanning numerous unknown semantic categories. To this end, we estimate (2) using Monte Carlo approximation (Kuhn et al., 2023; Farquhar et al., 2024):

$$\text{SE}(q) \approx -|\mathcal{C}|^{-1} \sum_{k=1}^{|\mathcal{C}|} \log p\left(\mathcal{C}_k \mid \{q; \mathcal{A}\}\right), \quad (3)$$

where $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_K\}$ denotes the semantic cluster. In practice, we sample N responses $\{r_1, \ldots, r_N\}$ and perform clustering based solely on the final answer segment of each r_i to obtain the set of semantic classes $\mathcal{C}$. The cluster probability in (3) is subsequently calculated as:

$$p(\mathcal{C}_k \mid \{q; \mathcal{A}\}) = \sum_{i=1}^{N} \mathbb{I}[r_i \in \mathcal{C}_k]\, p(r_i \mid \{q; \mathcal{A}\}). \quad (4)$$

It's worth noting that, since the thinking part of r_i can often span tens of thousands of tokens, we compute only the probability of the answer part in practical computation[2].

Based on the obtained SE for the response generated in round i, we design a mechanism to determine whether to terminate the reasoning process.

[2]This leads to inflated probability estimates, occasionally resulting in negative SE values. However, since these occurrences do not invalidate the observed negative correlation between SE and model performance, we do not apply corrections to the metric.

Specifically, leveraging the clear inverse correlation observed between SE and performance where higher SE signals greater uncertainty and lower response quality, we dynamically control the reasoning process by monitoring SE after each iteration. The reasoning continues into the next round if the measured SE exceeds a set threshold, implying the output requires further refinement, and is terminated only once the SE meets a pre-defined stopping condition. The specific configuration of this stopping condition will be described later.

Answer Selection Strategy. Once the reasoning process has terminated, we need to select the final response from the candidate answers. Without recourse to external validators, we employ three conventional strategies: random selection, maximum probability selection, and majority voting; the effectiveness of these approaches will be thoroughly examined in our experiments. Additionally, a worthy question is how to define the candidate answer pool. In this paper, instead of collecting all responses across the previous round, we propose to exclusively utilize responses generated in the terminal reasoning round as the candidate set. This design is motivated by our observation that the preceding output is likely to be suboptimal due to the higher elevated uncertainty. Incorporating such under-refined responses would introduce detrimental noise, thereby degrading model performance.

2.2 Adaptive Termination with Pre-defined Threshold

This section details the procedure for establishing reasoning termination conditions. A widely adopted approach is to establish a pre-defined SE threshold, halting the inference procedure when monitored SE is lower than this calibrated value. To this end, we first conducted foundational experiments using DeepSeek-R1-Distill-Qwen-7B model on our proprietary dataset of 1000 challenging mathematical problems. For each problem, we randomly selected one candidate solution from parallel inference outputs for evaluation. Then we analyzed the SE distribution of correct and incorrect answers, and the statistical analysis is visualized in Fig. 3. The statistical analysis reveals an $80/20$ pattern wherein about 80% of the correct answers lie in the lowest quintile of the SE distribution. Specifically, for $N = 2$, approximately 74% of correct answers fall within the specified range, increasing to 77% at $N = 4$ and reaching 78% for $N = 8$. Leveraging this pattern, given the model and a specific parallel degree, we can first sample parallel responses on substantial data. Then, we compute the empirical SE distribution and select the 20th-percentile value as the pre-defined threshold.

2.3 Adaptive Threshold-free Mechanism

Although the threshold-based approach described above provides a principled stopping criterion, it requires extensive pre-sampling and recalibration when model configurations or parallel settings change. We consequently develop a more adaptive threshold-free method to avoid this shortcoming. To this end, we reformulate our goal to identify the optimal reasoning round exhibiting minimal SE under a fixed inference budget. Surprisingly, we found that this goal is quite similar to the "Secretary Problem (Ferguson, 1989)". The classical secretary problem constitutes a foundational problem in Optimal Stopping Theory (Shiryaev, 1980), addressing sequential selection under uncertainty. It aims to maximize the probability of selecting the best candidate from an unknown sequence of applicants when interviews must be conducted irreversibly without recall. The core strategy used in secretary problem establishes a qualification baseline by observing the first T candidates during an initial exploration phase. Subsequent candidates are evaluated against this dynamically determined threshold, with immediate selection of the first applicant exceeding the baseline. This observation-then-selection framework aligns with our objective of identifying optimal termination round during multi-round reasoning. Drawing inspiration from this problem, we propose an adaptive threshold-free approach. Specifically, our method dynamically sets the adaptive threshold as the minimal SE observed during the initial T rounds and terminates inference immediately when encountering any round with SE below this calibrated minimum. Given the computational expense of long CoT sampling, we set $T = 1$, defining the dynamic threshold exclusively from the first reasoning round. This reduces computational overhead while maintaining empirically competitive performance.

3 Experiments

3.1 Experimental Setup

We evaluate the proposed SEAT method on AIME-2024, AIME-2025, MATH-500 (Hendrycks et al., 2021), MINERVA, and GPQA (Rein et al., 2023).

Method	AIME-2024		AIME-2025		MATH-500		MINERVA		GPQA	
	R1-7B	Qwen3-8B	R1-7B	Qwen3-8B	R1-7B	Qwen3-8B	R1-7B	Qwen3-8B	R1-7B	Qwen3-8B
Vanilla	60.41	75.33	37.50	69.33	93.95	95.25	52.14	59.20	53.09	59.03
Sequential Scaling										
S1	65.78	75.75	42.35	68.33	94.37	96.68	54.63	60.20	55.15	58.71
TT	67.08	78.00	42.50	72.17	94.70	96.00	54.57	59.29	55.25	60.86
Parallel Scaling										
MV	71.42	80.00	52.83	73.33	95.35	96.98	56.02	60.11	57.09	62.25
SMV	71.37	83.58	52.92	72.42	95.03	96.05	56.27	61.49	57.47	62.13
Hybrid Scaling										
LeaP†	64.38	–	41.25	–	—	–	—	–	55.56	–
Ours	**72.38**	**85.17**	**55.00**	**77.08**	**95.88**	**97.65**	**56.74**	**62.04**	**58.95**	**63.19**

Table 1: Performance comparison on various benchmarks. Results marked with † are from original papers.

Dataset	Qwen3-8B	R1-7B	R1-32B
AIME-24	80.42 / **85.17**	71.67 / **72.38**	84.00 / **85.83**
AIME-25	74.17 / **77.08**	53.67 / **55.00**	66.67 / **69.17**
MATH-500	97.20 / **97.65**	95.60 / **95.88**	**97.40** / 97.08
MINERVA	60.32 / **62.04**	56.15 / **56.74**	59.18 / **59.36**
GPQA	62.64 / **63.19**	57.56 / **58.95**	68.52 / **69.32**

Table 2: Performance comparison between our method and Majority Voting under the same inference budget constraints.

These benchmarks cover multiple domains and difficulty levels, allowing for a thorough assessment of reasoning performance across varied scenarios.

We compare the proposed SEAT with the following three groups of approaches. *1) Sequential Scaling Methods:* S1 (Muennighoff et al., 2025), Think twice, shortened as TT (Tian et al., 2025). *2) Parallel Scaling Approaches:* Majority Voting, denoted as MV (Wang et al., 2022), Shortest Majority Vote, shorted as SMV (Zeng et al., 2025). *3) Hybrid Scaling Methods:* LeaP (Luo et al., 2025). We use DeepSeek-R1-Distill-Qwen-7B (R1-7B)[3], DeepSeek-R1-Distill-Qwen-32B (R1-32B)[4], and Qwen3-8B[5] for our experimental analysis. For reproducibility, we set $M = 8$ and $N = 8$ for sequential[6] and parallel scaling methods, respectively. The maximum generation length is set to 32768 tokens, while temperature and top-p are applied to 0.7 and 0.95, respectively. Each query across all datasets will be repeated 8 times and the average accuracies are reported.

[3]https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

[4]https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B

[5]https://huggingface.co/Qwen/Qwen3-8B

[6]For method S1, we force the model to perform the next round of sequential reasoning by removing the end-of-thinking token and adding the string "Wait".

3.2 Performance Comparison

Tab. 1 compares our method with baselines on R1-7B and Qwen3-8B, where results labeled "Ours" utilize the SEAT framework with adaptive threshold-free termination. We observe that all scaling methods surpass the baseline, which validates the effectiveness of test-time scaling. Notably, parallel approaches demonstrate superior performance over sequential ones, even with the same maximum parallelism ($N = M = 8$). This can be attributed to the tendency of sequential methods to become trapped in loops of incorrect reasoning, whereas parallel methods avoid this issue through independent inference paths. By effectively leveraging the strengths of both strategies to enhance interaction between diverse reasoning paths, SEAT outperforms all comparative models, empirically validating the effectiveness. To ensure a fairer assessment of efficiency, Tab. 2 presents a controlled comparison against Majority Voting. Here, we allocate the exact same inference budget to the baseline by matching the specific number of LLM calls used by SEAT for each query. Under these identical constraints, our method still significantly outperforms Majority Voting across all datasets, particularly on Qwen3-8B, highlighting the efficiency of our adaptive resource allocation.

3.3 Analysis of Termination Strategies

We evaluate the effectiveness of different termination strategies as follows: ►*Fixed:* using the pre-defined threshold for inference termination. ►*Adaptive:* applying the adaptive threshold-free mechanism. ►*Min:* selecting the inference round with minimal SE among M round.

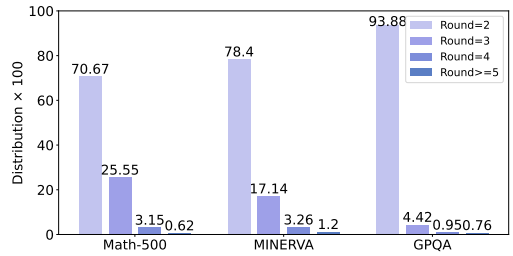
The experimental results with R1-7B and R1-32B as base models are shown in Tab. 3. Our ap-

Parallel Reasoning		AIME-2024		AIME-2025		MATH-500		MINERVA		GPQA	
		R1-32B	R1-7B	R1-32B	R1-7B	R1-32B	R1-7B	R1-32B	R1-7B	R1-32B	R1-7B
Random											
N=2	Baseline	70.83	60.41	53.33	37.50	95.12	93.95	57.36	52.14	66.87	53.09
	Ours (Fixed)	75.42	65.00	56.25	47.08	95.45	94.93	58.50	54.27	67.93	55.05
	Ours (Adaptive)	78.75	64.58	60.83	46.67	96.05	95.28	58.72	53.54	68.06	55.11
	Ours (Min)	**82.35**	**66.67**	**65.83**	**51.25**	**96.20**	**95.20**	**58.78**	**54.83**	**68.12**	**55.68**
N=4	Baseline	70.83	58.33	53.45	42.92	95.11	94.03	57.14	52.58	66.54	51.58
	Ours (Fixed)	78.33	68.75	64.17	46.25	96.03	95.03	**59.28**	**56.57**	67.23	55.18
	Ours (Adaptive)	80.41	68.33	65.25	48.17	96.38	95.50	58.51	55.97	67.42	55.47
	Ours (Min)	**83.33**	**70.83**	**70.83**	**48.75**	**96.65**	**95.80**	57.58	54.64	**67.80**	**56.25**
N=8	Baseline	70.42	56.25	52.92	40.83	95.13	93.78	57.85	52.80	65.85	53.09
	Ours (Fixed)	80.42	68.75	65.83	50.00	96.53	95.45	**59.56**	55.38	66.60	**57.95**
	Ours (Adaptive)	85.67	68.33	66.25	50.00	96.85	95.35	58.36	**55.71**	68.57	**57.95**
	Ours (Min)	**85.83**	**71.67**	**69.16**	**51.25**	**96.98**	**95.53**	58.55	55.00	**68.88**	57.07
Max Probability											
N=2	Baseline	72.92	64.17	53.73	41.67	95.22	94.37	58.01	53.08	66.74	54.37
	Ours (Fixed)	74.17	65.00	55.00	46.67	95.43	95.03	58.87	53.77	67.74	**55.81**
	Ours (Adaptive)	79.32	65.42	60.83	46.25	96.25	**95.35**	59.13	54.78	68.12	55.18
	Ours (Min)	**82.50**	**65.83**	**65.83**	**51.67**	**96.25**	95.28	**59.45**	**55.10**	**68.12**	55.43
N=4	Baseline	72.50	66.67	55.83	44.58	95.54	94.78	57.00	54.61	66.54	54.73
	Ours (Fixed)	75.83	70.42	59.17	48.75	95.98	95.60	58.60	55.28	66.67	55.05
	Ours (Adaptive)	81.67	70.00	67.08	50.00	**96.70**	**95.88**	**59.88**	**55.84**	67.42	56.00
	Ours (Min)	**83.33**	**71.67**	**71.25**	**52.08**	96.65	95.70	57.67	55.10	**67.42**	**56.44**
N=8	Baseline	70.83	64.58	55.00	45.42	95.33	94.55	57.58	54.87	65.66	56.07
	Ours (Fixed)	79.17	68.75	64.17	50.42	95.78	95.68	58.04	55.42	66.35	**58.84**
	Ours (Adaptive)	85.87	68.75	67.08	50.83	96.75	**95.95**	58.86	**56.27**	68.87	**58.84**
	Ours (Min)	**86.25**	**70.42**	**71.25**	**51.67**	**96.98**	95.60	**59.05**	55.23	**69.13**	57.95
Majority Voting											
N=2	Baseline	72.92	63.75	53.75	41.25	95.12	94.48	58.19	53.45	67.68	55.19
	Ours (Fixed)	74.17	65.00	55.00	46.25	95.45	95.00	58.95	54.04	67.87	**55.74**
	Ours (Adaptive)	80.00	**65.42**	61.67	46.67	96.30	**95.38**	59.35	**54.83**	68.06	55.68
	Ours (Min)	**82.50**	**65.42**	**65.83**	**51.25**	**96.23**	95.25	**59.45**	54.73	**68.08**	55.49
N=4	Baseline	80.83	70.33	64.17	48.58	95.84	95.80	58.47	55.02	67.84	55.78
	Ours (Fixed)	81.43	**72.08**	66.67	49.58	**96.80**	95.83	**60.07**	**55.79**	**68.24**	56.19
	Ours (Adaptive)	82.50	71.67	68.33	49.58	96.53	**95.93**	59.01	55.10	68.12	**56.44**
	Ours (Min)	**83.33**	70.83	**70.83**	**52.08**	96.67	95.83	57.76	55.00	67.55	56.19
N=8	Baseline	83.75	71.42	65.83	52.83	96.55	95.35	59.05	56.02	68.32	57.09
	Ours (Fixed)	85.83	**72.92**	70.42	52.92	96.98	**95.98**	**59.78**	56.42	**69.44**	58.84
	Ours (Adaptive)	85.83	72.38	69.17	**55.00**	**97.08**	95.88	59.36	**56.74**	69.32	**58.95**
	Ours (Min)	**85.83**	70.83	**70.83**	52.08	97.00	95.60	58.69	56.19	68.94	57.51

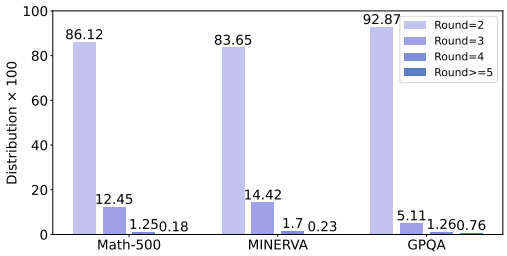
Table 3: The performance of variants of SEAT on different datasets. The baseline refers to the base models.

proach achieves substantial performance gains of both R1-32B and R1-7B models across different datasets. Specifically, under the random pick selection and adaptive setting, the R1-32B model achieves remarkable improvements, including an increase from 70.83 to 85.67 (+21.0%) on AIME-2024 and 53.33 to 66.25 (+24.2%) on AIME-2025, along with an average gain of of +2.5% across MATH-500, MINERVA, and GPQA. These results collectively confirm the effectiveness of our framework. Notably, even at minimal parallelization ($N = 2$), our method delivers remarkable gains with the R1-32B model rising from 53.33 to 60.83 (+14.1%) and the R1-7B model advancing from 37.50 to 46.67 (+24.5%) on AIME-2025. Furthermore, integrating max probability and majority voting strategies yields additional performance gains, with R1-32B model showing 0.6% (max probability) and 1.5% (majority voting) average improvements and R1-7B model achieving 1.1% and 4.0% gains, respectively. This demonstrates the scalability of our proposed framework.

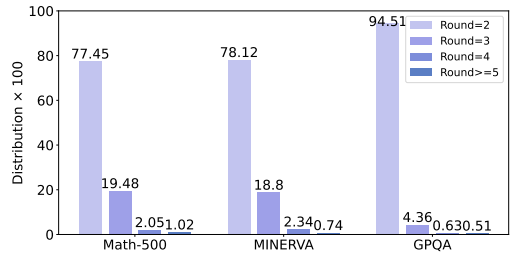
Regarding specific strategies, our approach using pre-defined thresholds significantly outperforms the baselines, demonstrating that pre-sampling probes of model SE distributions is an effective



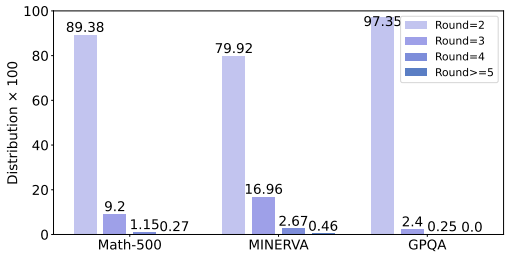
(a) R1-32B under $N = 2$ parallelization.



(b) R1-32B under $N = 8$ parallelization.



(c) R1-7B under $N = 2$ parallelization.



(d) R1-7B under $N = 8$ parallelization.

Figure 4: Round distribution across models and datasets during different parallel settings with adaptive termination.

strategy. Notably, although the thresholds were determined exclusively using mathematical datasets, they also improve performance on GPQA, demonstrating the robust generalization capability of the framework. Additionally, the adaptive threshold-free method usually surpasses fixed approaches. We attribute this to divergent SE distributions across different problem types for identical models. Establishing an SE baseline from the model's earlier reasoning outputs can better capture these dynamics, thereby yielding superior performance and improved generalization capability. Surprisingly, minimum selection (min) sometimes degrades performance in the R1-7B model at higher N values. For instance, at $N = 8$, the AIME-2025 score drops to 52.08 versus the baseline 52.83. We attribute this phenomenon to SE collapse during parallel inference in smaller models and will present a comprehensive visual analysis in Fig. 5.

3.4 Analysis of the Number of Inference Rounds

In this section, we analyze the inference round at which the model terminates generation. Fig. 4 illustrates the distribution of stopping rounds across various model scales, parallel configurations, and datasets. Note that termination occurs no earlier than round 2, as round 1 is reserved for evaluating the dynamic SE baseline. As shown in Fig. 4, over 70% of inferences terminate at the second round. This suggests that the model successfully leverages

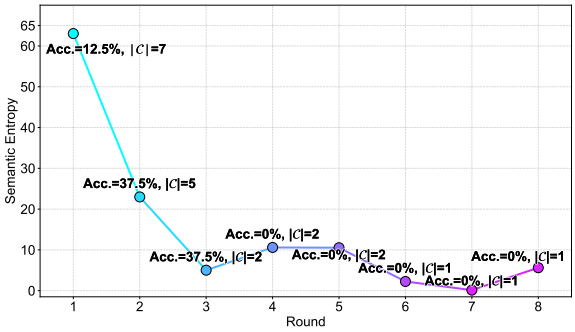


Figure 5: The evolution of SE and accuracy across different inference rounds of R1-7B model.

the output from the first round to refine subsequent responses and reduce uncertainty, as evidenced by the lower SE scores in round 2 compared to round 1. Furthermore, our proposed adaptive termination mechanism finishes inference within 3 rounds in most scenarios, thereby avoiding the substantial computational cost $O(N \cdot M)$ of running all sequential steps. Moreover, scaling model size or increasing the parallelization degree (higher N) amplifies second-round termination rates. This scaling behavior confirms that greater model capacity or expanded parallel search reinforces self-refinement capabilities, yielding higher-quality responses through earlier convergence.

3.5 Visualization of SE Collapse

We investigate an anomaly observed in the R1-7B model ($N = 8$), where the minimum SE selection

yielded lower accuracy than the baseline. To probe this, we analyze a representative case, tracking the evolution of SE and $\mathbb{E}(Acc.)$. As depicted in Fig. 5, while SE plummets from 63.04 to 0.14 by round-7, $\mathbb{E}(Acc.)$ briefly peaks at 37.5 before collapsing to 0. Concurrently, the diminishing number of semantic clusters ($|\mathcal{C}|$) throughout the inference process signals a loss of reasoning diversity, driving the model toward overconfident errors. Upon examining representative outputs at round-7, we notice a remarkable reduction in response length. The model often skips detailed reasoning steps, directly outputting concise answers (specific examples are provided in the Appendix A). We term this degradation "semantic entropy collapse", a precipitous SE deterioration resulting in vanishing diversity and blind generation. We attribute this phenomenon to the limited reasoning power of smaller models (e.g., 7B), as it remains absent in R1-32B. Notably, in this case, our proposed adaptive termination strategy intervenes at round-2, securing the highest expected accuracy and averting SE collapse to maintain the small models' performance.

4 Related Works

Test-time scaling allows large language models (LLMs) to engage in more deliberative reasoning before producing final answers. Recent research highlights that scaling test-time compute can often be more effective than scaling model parameters (Snell et al., 2024; Wu et al., 2024), provided that computational resources are allocated optimally (Wang et al., 2025b). Existing approaches to leverage this compute can be broadly categorized into three main strategies, as discussed below.

Parallel Scaling. Parallel scaling prompts LLMs to independently generate multiple outputs, with the final answer typically selected by unsupervised selection methods such as majority voting (Wang et al., 2022; Chen et al., 2023). To address the cost of fixed-sample voting, Self-Consistency variants have been proposed. Early-stopping mechanisms improve efficiency by dynamically halting generation once a reliable consensus is reached (Aggarwal et al., 2023; Li et al., 2024). Additionally, fine-grained methods integrate segment-level commonalities, extending parallel scaling to free-form generation (Wang et al., 2024). While incorporating external verifiers can further improve selection quality (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024; Brown et al., 2024), standard parallel scaling generally lacks coordination between samples, limiting its support for iterative refinement.

Sequential Scaling. Sequential scaling allows LLMs to perform extended chain-of-thought reasoning, incorporating behaviors like verification, backtracking, and subgoal decomposition (Gandhi et al., 2025). This method has advanced complex reasoning in models such as OpenAI O1 (Jaech et al., 2024) and DeepSeek R1 (DeepSeek-AI et al., 2025). Control signals (e.g., "wait," "Final answer") are used to manage computational budget (Muennighoff et al., 2025; Zhang et al., 2025). Multi-round strategies further refine answers by iteratively feeding previous outputs back into the model (Tian et al., 2025). While sequential scaling supports more deliberate reasoning, LLMs can still become stuck in incorrect reasoning paths, struggling to recover correct answers (Zeng et al., 2025; Luo et al., 2025).

Hybrid Scaling. Recent work explores hybrid strategies combining parallel exploration with sequential conditioning. Pan et al. (2025) propose adaptive parallel reasoning, where parent threads decompose tasks for child threads and aggregate their results. Similarly, Luo et al. (2025) introduces a routing mechanism that enables information exchange among multiple independent chain-of-thought processes for iterative refinement. Unlike prior intervention-heavy methods, we initiate flexible parallel trajectories to guide refinement, utilizing unsupervised semantic entropy to dynamically allocate reasoning steps based on task complexity.

5 Conclusion

This work proposes SEAT, a novel test-time scaling framework that utilizes semantic entropy to synergize parallel exploration and sequential refinement without the need for training. We demonstrate the critical role of semantic entropy in gauging parallel reasoning quality, allowing SEAT to dynamically adapt the depth of sequential reasoning. Extensive evaluations across five challenging benchmarks confirm that SEAT significantly enhances reasoning performance. Furthermore, its dynamic termination strategy effectively mitigates the semantic entropy collapse often observed in compact models, ensuring robust multi-round reasoning.

Limitations

Despite the promising results, our work has several limitations. First, our experiments primarily focus on tasks with definitive answers, such as mathematics and scientific QA, where semantic entropy is straightforward to compute via answer clustering. The applicability of our method to open-ended generation tasks (e.g., creative writing), where answer equivalence is harder to define, remains to be explored. Second, regarding the adaptive strategy, we set $T = 1$ to establish the baseline primarily for computational efficiency. While effective, exploring more flexible or dynamic baseline update strategies could potentially yield further performance improvements. Ultimately, our evaluation primarily focuses on the accuracy of the final answer. We have not conducted a fine-grained assessment of the reasoning process itself, including aspects such as logical coherence and interpretability, which remains an important direction for future research.

Ethics Statement

This study strictly adheres to the community ethical guidelines. The datasets utilized in this work are publicly available benchmarks focused on mathematical and scientific reasoning. These datasets are devoid of personally identifiable information or discriminatory content. Furthermore, we strictly comply with the licenses of all models and datasets employed in this research. Therefore, we foresee no direct negative societal impacts from this work.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and 1 others. 2023. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms. *arXiv preprint arXiv:2305.11860*.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. 2025. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *ArXiv*, abs/2505.12346.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, and 1 others. 2024. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation. *ArXiv*, abs/2311.17311.

Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625 – 630.

Thomas S. Ferguson. 1989. Who solved the secretary problem. *Statistical Science*, 4:282–289.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Xiaodong Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS Datasets and Benchmarks*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Zhewei Kang, Xuandong Zhao, and Dawn Xiaodong Song. 2025. Scalable best-of-n selection for large language models via self-certainty. *ArXiv*, abs/2502.18581.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ArXiv*, abs/2302.09664.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. 2024. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. *arXiv preprint arXiv:2401.10480*.

Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Feiyu Xiong, and Zhiyu Li. 2024. Internal consistency and self-feedback in large language models: A survey. *ArXiv*, abs/2407.14507.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Tongxu Luo, Wenyu Du, Jiaxi Bi, Stephen Chung, Zhengyang Tang, Hao Yang, Min Zhang, and Benyou Wang. 2025. Learning from peers in reasoning models. *ArXiv*, abs/2505.07787.

Andrey Malinin and Mark John Francis Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Fei-Fei Li, Hanna Hajishirzi, Luke S. Zettlemoyer, Percy Liang, Emmanuel J. Candes, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *ArXiv*, abs/2501.19393.

OpenAI. 2024. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/.

Jiayi Pan, Xiuyu Li, Long Lian, Charlie Snell, Yifei Zhou, Adam Yala, Trevor Darrell, Kurt Keutzer, and Alane Suhr. 2025. Learning adaptive parallel reasoning with language models. *ArXiv*, abs/2504.15466.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *ArXiv*, abs/2311.12022.

Albert N. Shiryaev. 1980. Optimal stopping rules. In *International Encyclopedia of Statistical Science*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b/.

Xiaoyu Tian, Sitong Zhao, Haotian Wang, Shuaiting Chen, Yunjie Ji, Yiping Peng, Han Zhao, and Xiangang Li. 2025. Think twice: Enhancing llm reasoning by scaling multi-round test-time thinking. *ArXiv*, abs/2503.19855.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025a. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning.

Xinglin Wang, Yiwei Li, Shaoxiong Feng, Peiwen Yuan, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. 2024. Integrate the essence and eliminate the dross: Fine-grained self-consistency for free-form language generation. *arXiv preprint arXiv:2407.02056*.

Xinglin Wang, Yiwei Li, Shaoxiong Feng, Peiwen Yuan, Yueqi Zhang, Jiayi Shi, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. 2025b. Every rollout counts: Optimal resource allocation for efficient test-time scaling. *arXiv preprint arXiv:2506.15707*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, and 1 others. 2025c. Thoughts are all over the place: On the underthinking of o1-like llms. *arXiv preprint arXiv:2501.18585*.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? *ArXiv*, abs/2502.12215.

Junyu Zhang, Runpei Dong, Han Wang, Xuying Ning, Haoran Geng, Peihao Li, Xialin He, Yutong Bai, Jitendra Malik, Saurabh Gupta, and Huan Zhang. 2025. Alphaone: Reasoning models thinking slow and fast at test time. *arXiv preprint arXiv:2505.24863*.

A Example Outputs for Semantic Entropy Collapse

The phenomenon of semantic entropy collapse can be seen through the following three illustrative examples.

Example 1

...<\think>
The inradius of triangle ABC is 35, expressed as the fraction 35/1. Thus, m + n equals 35 + 1, which is 36.
[
\boxed36
]

Example 2

...<\think>
The inradius of triangle ABC is the sum of the radii of the two tangent circles, which are 34 and 1. Therefore, the inradius $r = 34+1 = 35$. Expressed as a fraction, this is $\frac{35}{1}$. The sum of the numerator and denominator is $35 + 1 = 36$.
[
\boxed36
]

Example 3

...<\think>
The inradius of triangle ABC is the sum of the radii of the two tangent circles, which are 34 and 1. Therefore, the inradius $r = 34+1 = 35$. Expressed as a fraction, this is $\frac{35}{1}$, so $m = 35$ and $n = 1$. Thus, $m + n = 35 + 1 = 36$.
[
\boxed36
]