ARISE: An Adaptive Resolution-Aware Metric for Test-Time Scaling Evaluation in Large Reasoning Models

Zhangyue Yin[♢] **Qiushi Sun**[♡] **Zhiyuan Zeng**[♢] **Zhiyuan Yu**[♠]
Qipeng Guo[♣☾] **Xuanjing Huang**[♢*] **Xipeng Qiu**[♢☾*]
[♢]Fudan University [♡]The University of Hong Kong
[♠]Nanjing University [♣]Shanghai AI Laboratory [☾]Shanghai Innovation Institute
{yinzy21,cengzy23}@m.fudan.edu.cn qiushisun@connect.hku.hk
zhiyuan_yu@smail.nju.edu.cn guoqipeng@pjlab.org.cn
{xpqiu,xjhuang}@fudan.edu.cn

Abstract

Test-time scaling has emerged as a transformative paradigm for enhancing the performance of large reasoning models, enabling dynamic allocation of computational resources during inference. However, as the landscape of reasoning models rapidly expands, a critical question remains: how can we systematically compare and evaluate the test-time scaling capabilities across different models? In this paper, we introduce ARISE (Adaptive Resolution-aware Inference Scaling Evaluation), a novel metric specifically designed to assess the test-time scaling effectiveness of large reasoning models. Unlike existing evaluation approaches, ARISE incorporates two key innovations: (1) sample-level awareness that effectively penalizes negative scaling behaviors where increased computation leads to performance degradation, and (2) a dynamic sampling mechanism that mitigates the impact of accuracy fluctuations and token count instability on the final assessment. We conduct comprehensive experiments evaluating state-of-the-art reasoning models across diverse domains including mathematical reasoning, code generation, and agentic tasks. Our results demonstrate that ARISE provides a reliable and fine-grained measurement of test-time scaling capabilities, revealing significant variations in scaling efficiency across models. Notably, our evaluation identifies Claude Opus as exhibiting superior scaling characteristics compared to other contemporary reasoning models.

1 Introduction

Test-time scaling has emerged as a transformative paradigm in Large Reasoning Models, enabling dynamic computational resource allocation during inference to enhance model performance (Snell et al., 2025; Wu et al., 2025; Sun et al., 2025). As an increasing number of models with test-time scaling capabilities are released (Jaech et al., 2024; OpenAI, 2025a; DeepSeek-AI, 2025; Yang et al., 2025), the ability to scale effectively at inference has become a critical dimension for evaluating model capabilities alongside traditional metrics (Zhang et al., 2025; Zeng et al., 2024). However, systematically comparing test-time scaling effectiveness across diverse models presents significant methodological challenges.

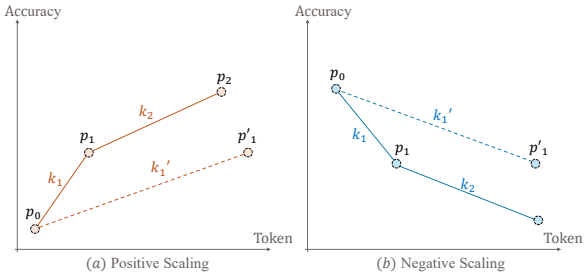

Figure 1: Limitations of slope-based metrics in test-time scaling evaluation. (a) When performance improves from p_0 to p_1 and p'_1, the steeper slope correctly rewards p_1 for achieving the same accuracy with fewer tokens. (b) When performance degrades, the slope metric incorrectly assigns a higher value to p'_1 despite it wasting more tokens for worse performance.

While scaling curves provide intuitive visualization of test-time scaling behavior, they lack the quantitative precision necessary for rigorous model comparison. Recent work by Muennighoff et al. (2025) proposed using slope-based metrics to quantify scaling capabilities. However, this approach entails two notable limitations. First, it operates at the aggregate accuracy level, overlooking a core objective of test-time scaling: *converting previously incorrect samples to correct ones* (Chen et al., 2024b). This aggregate view fails to penalize samples that become incorrect after scaling. Second, as illustrated in Figure 1, slope metrics exhibit pathological behavior under negative scaling scenarios. When performance improves (Figure 1a), the metric correctly rewards models that achieve higher accuracy with fewer tokens. However, when performance degrades (Figure 1b), the same metric paradoxically assigns higher scores to models that

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 5828–5848

waste more tokens while achieving worse results, a clear misalignment with the intended objective.

Beyond these conceptual limitations, fair measurement of test-time scaling capabilities faces practical challenges. Large reasoning models typically require high sampling temperatures to explore diverse solution paths (Yang et al., 2025), introducing substantial variability in both token consumption and accuracy measurements. This inherent stochasticity makes it difficult to obtain stable, reproducible assessments of scaling behavior across different models and evaluation runs.

To address these challenges, we introduce ARISE (**A**daptive **R**esolution-aware **I**nference **S**caling **E**valuation), a novel metric specifically designed for robust evaluation of test-time scaling in large reasoning models. ARISE incorporates two key innovations: (1) **sample-level awareness** that tracks individual sample trajectories across scaling levels, effectively penalizing both samples that degrade after scaling and wasteful token consumption under negative scaling; and (2) **dynamic sampling mechanism** that adaptively adjusts sample-level evaluation runs based on observed variance in accuracy and token consumption, ensuring statistically reliable measurements. These design choices make ARISE a principled and stable metric for assessing test-time scaling capabilities.

We conduct comprehensive experiments evaluating state-of-the-art reasoning models across diverse domains, including mathematical reasoning, code generation, and agentic tasks. Our empirical analysis demonstrates that ARISE delivers consistent and fine-grained measurements of test-time scaling effectiveness, exposing substantial disparities in scaling efficiency across models. Notably, Claude Opus outperforms all contemporary models, attaining the highest ARISE scores while exhibiting robust stability across all evaluated task domains. Our contributions are summarized as follows:

- We introduce scaling efficiency as a critical dimension for evaluating reasoning model abilities and identify fundamental limitations in existing test-time scaling evaluation methods.
- We propose ARISE, a novel evaluation metric that incorporates sample-level awareness and dynamic sampling mechanism to ensure statistically reliable measurements.
- We present a comprehensive empirical evaluation across multiple domains, establishing ARISE as a reliable metric for comparing test-time scaling capabilities of reasoning models.

2 Related Work

Test-Time Scaling As training-time scaling approaches its computational and data limits (Villalobos et al., 2024), test-time scaling has emerged as a promising new paradigm for advancing model capabilities (Zhang et al., 2025; Zeng et al., 2025; Wu, 2025). Recent empirical studies have demonstrated that optimal test-time compute allocation can be more effective than simply scaling model parameters (Snell et al., 2025; Wu et al., 2025). The release of OpenAI's o1 model (Jaech et al., 2024) has catalyzed a surge of research into understanding and improving test-time scaling mechanisms (Chen et al., 2024b; Hu et al., 2025; Luo et al., 2025b), with process reward models further enabling fine-grained guidance of reasoning steps during inference (Lightman et al., 2024; Yin et al., 2025). Major models including OpenAI GPT-5 (OpenAI, 2025a), Anthropic Claude (Anthropic, 2025b), DeepSeek-V3.1 (DeepSeek-AI, 2024) have successfully integrated test-time scaling capabilities. Test-time scaling has been widely applied to a diverse range of complex tasks including mathematical reasoning (Wang et al., 2025a; Balachandran et al., 2025), code generation (Yu et al., 2025; Li et al., 2025; Sun et al., 2024), and agentic tasks such as automated scholarly novelty assessment (Zhang et al., 2026; Zhu et al., 2025; Chakraborty et al., 2025).

Scaling Evaluation The evaluation of test-time scaling encompasses multiple dimensions that capture different aspects of model behavior. From a performance perspective, Pass@1 remains the most prevalent metric (Yang et al., 2025), serving as the standard benchmark for mathematical reasoning (Team, 2025; Hendrycks et al., 2021) and code generation tasks (Raihan et al., 2024; Jain et al., 2025). Extensions such as Pass@k and Cons@k (Chen et al., 2021) provide models with multiple attempts, offering a more comprehensive view of their problem-solving capabilities (Brown et al., 2024; Li et al., 2022). Beyond static task-level accuracy, recent efforts have explored richer evaluation paradigms that capture dynamic aspects of model behavior, such as sequential problem-solving settings that quantify both learning capability and efficiency over successive attempts (Dou et al.) and context learning benchmarks that as-

sess a model's ability to acquire and apply novel knowledge presented at inference time (Dou et al., 2026). More broadly, the methodology of large language model evaluation itself has received growing attention, with systematic studies investigating scoring standards, evaluator types, and ranking systems to improve the reliability and consistency of evaluation outcomes (Zhang et al., 2024, 2023). From an efficiency standpoint, recent work has examined computational overhead through various lenses, including token consumption and step redundancy (Luo et al., 2025a; Chiang and Lee, 2024; Chen et al., 2024a). Notably, Wang et al. (2025b) identified a critical phenomenon where reasoning models exhibit excessive trajectory switching during inference, leading to insufficient depth of exploration, a behavior they quantify through the proposed Underthinking Score. Another crucial dimension concerns controllability, the ability of models to consistently scale their reasoning process to predetermined computational budgets (Bhargava et al., 2023; Muennighoff et al., 2025; Aggarwal and Welleck, 2025).

The most comprehensive approach to scaling evaluation employs scaling curves (Wu et al., 2025; Teng et al., 2025), which simultaneously capture both accuracy improvements and computational efficiency. These curves visualize the trade-off between performance gains and resource utilization, providing insights into the marginal utility of additional computation. To quantify this relationship, the scaling metric (Muennighoff et al., 2025) computes the average gradient across all point pairs on the curve:

$$\text{Scaling} = \frac{1}{\binom{|\mathcal{P}|}{2}} \sum_{\substack{p_1,p_2\in\mathcal{P} \\ \mathcal{T}(p_2)>\mathcal{T}(p_1)}} \frac{\mathcal{A}(p_2)-\mathcal{A}(p_1)}{\mathcal{T}(p_2)-\mathcal{T}(p_1)} \quad (1)$$

where $\mathcal{P}$ represents the set of points on the scaling curve, and functions $\mathcal{A}(\cdot)$ and $\mathcal{T}(\cdot)$ denote the accuracy and token consumption at each point, respectively. This formulation provides a scalar measure of scaling efficiency but fails to capture sample-level variations and negative scaling behaviors that are crucial for effective evaluation.

3 ARISE: Adaptive Resolution-aware Inference Scaling Evaluation

We propose ARISE, a novel metric that addresses the limitations of existing test-time scaling evaluation approaches through sample-level error awareness and dynamic sampling mechanisms.

3.1 Metric Design

For each sample i in the evaluation dataset, we define $a_i^{(j)} \in \{0,1\}$ as the binary accuracy at scaling iteration j, and $t_i^{(j)}$ as the corresponding token consumption. The ARISE score for sample i is computed as:

$$\text{ARISE}_i = \sum_{j=1}^{m} \Delta a_i^{(j)} \cdot W_i^{(j)} \quad (2)$$

where $\Delta a_i^{(j)} = a_i^{(j)} - a_i^{(j-1)}$ represents the accuracy change, and the weight function $W_i^{(j)}$ is defined as:

$$W_i^{(j)} = \left(\frac{t_i^{(j-1)}}{t_i^{(j)}}\right)^{\text{sign}(\Delta a_i^{(j)})} \quad (3)$$

The overall ARISE score aggregates individual sample scores:

$$\text{ARISE} = \frac{1}{n}\sum_{i=1}^{n} \text{ARISE}_i \quad (4)$$

Sample-Level Awareness. ARISE evaluates scaling behavior at the granularity of individual samples by examining transitions between adjacent scaling iterations. For any pair of consecutive iterations $(j-1, j)$, the contribution to ARISE_i depends on the accuracy transition:

$$C_i^{(j)} = \begin{cases} 0 & \text{if } a_i^{(j)} = a_i^{(j-1)} \\ \frac{t_i^{(j-1)}}{t_i^{(j)}} & \text{if } a_i^{(j)} = 1, a_i^{(j-1)} = 0 \\ -\frac{t_i^{(j)}}{t_i^{(j-1)}} & \text{if } a_i^{(j)} = 0, a_i^{(j-1)} = 1 \end{cases} \quad (5)$$

This formulation ensures that ARISE captures the critical moment when a sample transitions from incorrect to correct, while penalizing degradations where additional computation leads to errors. Since $t_i^{(j)} > t_i^{(j-1)}$ by construction, the penalty magnitude $|\frac{t_i^{(j)}}{t_i^{(j-1)}}| > |\frac{t_i^{(j-1)}}{t_i^{(j)}}|$ exceeds the reward magnitude, reflecting the asymmetric cost of computational waste.

Negative Scaling Correction. When performance deteriorates ($\Delta a_i^{(j)} < 0$), the sign function in Equation 3 becomes -1, transforming the weight to:

$$W_i^{(j)} = \left(\frac{t_i^{(j-1)}}{t_i^{(j)}}\right)^{-1} = \frac{t_i^{(j)}}{t_i^{(j-1)}} > 1 \quad (6)$$

This design amplifies penalties proportionally to token waste, which uses more computational resources for worse results receives progressively stronger penalties, directly addressing the fundamental limitation of existing metrics that fail to adequately penalize negative scaling behaviors.

Magnitude-Aware Design. Unlike conventional scaling metrics that employ absolute differences, ARISE utilizes ratios to enable relative scaling measurement adapted to problem difficulty. Consider a fixed token increment $\Delta t = 1000$:

For simple problems where $t_i^{(j-1)} = 1000$, an additional 1000 tokens doubles the computational budget, representing substantial additional reasoning capacity. Conversely, for complex problems where $t_i^{(j-1)} = 10000$, the same increment represents only a 10% increase, likely insufficient for meaningful additional analysis. This ratio-based approach ensures that scaling effectiveness is measured relative to the baseline computational requirements, providing more accurate assessments across problems of varying complexity.

Non-Combinatorial Computation. ARISE employs adjacent-pair computation rather than exhaustive pairwise combinations, avoiding redundancy and computational complexity. Consider a sequence of four scaling iterations with accuracy pattern $(0, 1, 0, 1)$ and strictly increasing tokens $t_i^{(0)} < t_i^{(1)} < t_i^{(2)} < t_i^{(3)}$.

The adjacent-pair approach yields:

$$\text{ARISE}_i = \frac{t_i^{(0)}}{t_i^{(1)}} - \frac{t_i^{(2)}}{t_i^{(1)}} + \frac{t_i^{(2)}}{t_i^{(3)}} \tag{7}$$

whereas combinatorial computation would include an additional term $\frac{t_i^{(0)}}{t_i^{(3)}}$, inappropriately rewarding the direct transition from initial failure to final success while ignoring the intermediate regression. This spurious reward could exceed the penalty for the intermediate failure, demonstrating why adjacent-pair computation provides more reasonable scaling assessment.

Boundedness Properties. Unlike traditional scaling metrics with symmetric bounds, ARISE exhibits asymmetric bounds reflecting its design philosophy. When $a_i^{(j-1)} = 0$ and $a_i^{(j)} = 1$, as $t_i^{(j)} \to t_i^{(j-1)}$, $\text{ARISE}_i \to 1^-$. For degradation cases where $a_i^{(j-1)} = 1$ and $a_i^{(j)} = 0$:

$$C_i^{(j)} = -\frac{t_i^{(j)}}{t_i^{(j-1)}} < -1 \tag{8}$$

Thus, ARISE $\in (-\infty, 1)$, though practical values typically remain within $(-1, 1)$ as extreme negative scaling is rare. The unbounded negative range ensures severe penalties for egregious computational waste, while the bounded positive range prevents over-rewarding improvements. We provide formal boundedness analysis in Appendix A.

3.2 Adaptive Sampling Strategy

Test-time scaling evaluation faces inherent variance in both accuracy outcomes and token consumption across trials. We introduce an adaptive sampling mechanism that dynamically allocates computational budget based on observed variance patterns, enhancing evaluation reliability.

Variance Characterization. For each sample i at scaling iteration j, we conduct an initial probing phase with m_{min} trials. Let $a_{i,k}^{(j)}$ and $t_{i,k}^{(j)}$ denote the accuracy and token consumption for trial k. We compute the empirical statistics:

$$\mu_{a_i^{(j)}} = \frac{1}{m_{\text{min}}} \sum_{k=1}^{m_{\text{min}}} a_{i,k}^{(j)} \tag{9}$$

$$\sigma_{a_i^{(j)}} = \sqrt{\frac{1}{m_{\text{min}}} \sum_{k=1}^{m_{\text{min}}} (a_{i,k}^{(j)} - \mu_{a_i^{(j)}})^2} \tag{10}$$

with analogous definitions for token statistics $\mu_{t_i^{(j)}}$ and $\sigma_{t_i^{(j)}}$.

Normalized Variance Measure. To enable fair comparison across different scales and magnitudes, we employ the coefficient of variation (CV):

$$\text{CV}_{a_i^{(j)}} = \frac{\sigma_{a_i^{(j)}}}{\mu_{a_i^{(j)}} + \epsilon} \tag{11}$$

$$\text{CV}_{t_i^{(j)}} = \frac{\sigma_{t_i^{(j)}}}{\mu_{t_i^{(j)}} + \epsilon} \tag{12}$$

where $\epsilon = 10^{-8}$ prevents division by zero. The combined variance indicator captures both dimensions:

$$\text{CV}_i^{(j)} = \text{CV}_{a_i^{(j)}} + \text{CV}_{t_i^{(j)}} \tag{13}$$

Model	AIME		HMMT		GPQA Diamond		MMLU-Pro	
	ARISE	SM×1000	ARISE	SM×1000	ARISE	SM×1000	ARISE	SM×1000
o1	0.1346	0.0607	0.1277	0.0183	0.1228	0.0385	0.1509	0.0504
o3	0.2993	0.0782	0.1673	0.0186	0.2124	0.0589	0.2789	0.0567
o3-mini	0.1306	0.0374	0.1888	0.0302	0.1649	0.0221	0.1663	0.0416
o4-mini	0.2402	0.0348	0.1673	0.0435	0.1994	0.0423	0.2080	0.0392
gpt-oss-20B	-0.4030	0.0205	-0.3126	0.0168	-0.3274	0.0224	-0.2694	0.0227
gpt-oss-120B	-0.3340	0.0273	-0.1999	0.0241	-0.2734	0.0286	-0.1615	0.0314
gpt-5	0.1566	0.0259	0.2996	0.0265	0.2185	0.0263	0.3186	0.0298
Claude Sonnet 4	0.1041	0.0461	0.0404	0.0109	0.0636	0.0326	0.1059	0.0264
Claude Opus 4	0.3475	0.0653	0.1717	0.0617	0.2212	0.0488	0.3330	0.0675
Claude Opus 4.1	**0.4529**	**0.1462**	**0.4709**	**0.1419**	**0.4454**	**0.1416**	**0.4932**	**0.2038**
Qwen-3-0.6B	0.2936	0.0023	0.1769	0.0071	0.2114	0.0052	0.2716	0.0034
Qwen-3-1.7B	0.3658	0.0278	0.2746	0.0256	0.3237	0.0321	0.3917	0.0345
Qwen-3-4B	0.2166	0.0496	0.2217	0.0378	0.2400	0.0325	0.2569	0.0619
Qwen-3-8B	0.3085	0.0342	0.3010	0.0268	0.3022	0.0354	0.3274	0.0357
Qwen-3-14B	0.3247	0.0684	0.2213	0.0473	0.2594	0.0449	0.3120	0.0801
Qwen-3-32B	0.3883	0.0767	0.2038	0.0585	0.2644	0.0576	0.3992	0.0658
Qwen3-30B-A3B	0.3293	0.0503	0.3855	0.0742	0.3727	0.0437	0.4162	0.0691
Qwen3-235B-A22B	0.3915	0.0819	0.4306	0.0415	0.4069	0.0762	0.4533	0.0794
Deepseek-R1	-0.0318	0.0072	-0.0455	0.0046	-0.0493	0.0033	-0.0108	0.0028
V3.1	0.3966	0.0351	0.2048	0.0369	0.2714	0.0427	0.3559	0.0362
V3.1-Terminus	0.3241	0.0237	0.2835	0.0353	0.3091	0.0316	0.3228	0.0319
V3.2-Exp	0.3029	0.0276	0.2651	0.0331	0.2726	0.0209	0.3219	0.0293

Table 1: Performance of mainstream models in mathematical and scientific reasoning. Each benchmark shows ARISE scores and corresponding Scaling Metrics (SM). For improved readability, SM values have been multiplied by 1000. The original unscaled values can be found in Appendix Table 4.

Dynamic Sampling Protocol. We implement an adaptive sampling strategy with a maximum sampling budget $m_{\max}$ and a convergence threshold τ. For each configuration (i, j), we iteratively collect samples while monitoring the combined coefficient of variation. Sampling continues until either convergence is achieved or the budget is exhausted:

$$\mathrm{CV}_i^{(j)} < \tau \quad \text{or} \quad k = m_{\max} \tag{14}$$

Upon termination at iteration k^*, we compute the final statistics as the empirical means over all collected samples:

$$a_i^{(j)} = \frac{1}{k^*} \sum_{k=1}^{k^*} a_{i,k}^{(j)} \tag{15}$$

$$t_i^{(j)} = \frac{1}{k^*} \sum_{k=1}^{k^*} t_{i,k}^{(j)} \tag{16}$$

This adaptive approach balances statistical reliability with computational efficiency. High-variance configurations receive additional sampling to reduce uncertainty, while stable configurations terminate early to conserve resources. We provide the algorithmic implementation in Appendix B.

Budget Allocation. For comparative analysis with fixed total budget B across n samples and J iterations, we allocate additional trials proportionally to observed variance:

$$m_i^{(j)} = m_{\min} + \left\lfloor \frac{(B - nJm_{\min}) \cdot \mathrm{CV}_i^{(j)}}{\sum_{i',j'} \mathrm{CV}_{i'}^{(j')}} \right\rfloor \tag{17}$$

This variance-proportional allocation distributes the remaining sampling budget according to the variability observed in the initial probing phase. Cases with greater fluctuations receive more sampling, ensuring a more robust and reliable assessment of scaling across diverse model behaviors.

4 Experiments

We conduct comprehensive experiments to evaluate ARISE across diverse reasoning tasks and model architectures. Our evaluation encompasses both text-based and multimodal benchmarks, covering mathematical reasoning, scientific problem-solving, code generation, and agentic capabilities.

4.1 Experimental Settings

Evaluation Datasets. To thoroughly assess the effectiveness of ARISE, we incorporate a comprehensive suite of text-based and multimodal benchmarks. Our evaluation spans four primary categories of reasoning tasks:

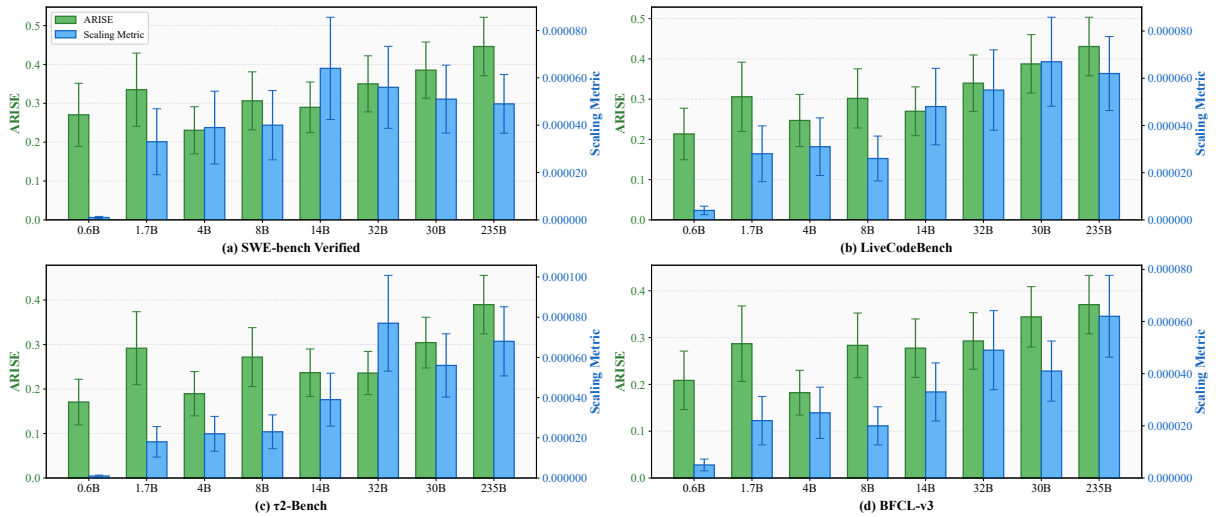

Figure 2: Comparison of ARISE and Scaling Metric across different Qwen3 models on code and agentic tasks. The x-axis shows model parameter counts where 0.6B, 1.7B, 4B, 8B, 14B, 32B correspond to Qwen-3 models, 30B corresponds to Qwen3-30B-A3B, and 235B corresponds to Qwen3-235B-A22B. ARISE values are shown on the left y-axis (green bars) while Scaling Metric values are shown on the right y-axis (blue bars). Error bars represent standard deviations across five independent runs. Complete results for all models are presented in Appendix Table 6.

- **Mathematical Reasoning:** AIME (Team, 2025) and HMMT (Balunović et al., 2025)
- **Scientific Reasoning:** GPQA-diamond (Rein et al., 2024) and MMLU-Pro (Wang et al., 2024a)
- **Code Generation:** SWE-bench Verified (Jimenez et al., 2024) and LiveCodeBench (Jain et al., 2025)
- **Agentic Tasks:** τ^2-Bench (Barres et al., 2025) and BFCL-v3 (Patil et al., 2025)

Additionally, we evaluate multimodal reasoning capabilities on vision-language tasks, including MMMU (Yue et al., 2024), MathVista (Lu et al., 2024), and CharXiv-Reasoning (Wang et al., 2024b). These benchmarks test the models' ability to integrate visual and textual information for complex reasoning. Detailed descriptions of each dataset, including sample counts and evaluation metrics, are provided in Appendix D.

Implementation Details. We evaluate both proprietary and open-source reasoning models through their respective interfaces. For the main experiments, we configure the adaptive sampling parameters with $m_{\min} = 3$, $m_{\max} = 10$, and convergence threshold $\tau = 0.5$. To assess metric stability (Section 4.3), we conduct five independent runs and report standard deviations. For the adaptive sampling analysis under fixed budget constraints (Section 4.4), we set the total budget $B = 5nJ$, where n denotes the number of samples and J represents the number of scaling iterations. Additional implementation details are documented in Appendix C.

4.2 Main Results

Mathematical and Scientific Reasoning. Table 1 presents our evaluation of models on mathematical and scientific reasoning benchmarks. Claude Opus 4.1 consistently achieves the highest performance on both ARISE and traditional Scaling Metric, with ARISE scores consistently exceeding 0.45 across most benchmarks and reaching 0.493 on MMLU-Pro. Among open-source models, Qwen3-235B-A22B demonstrates the strongest test-time scaling capabilities, achieving ARISE scores above 0.39 on all evaluated tasks, approaching commercial models. Notably, the ARISE metric reveals nuanced scaling behaviors that traditional metrics miss, particularly the negative scaling phenomena exhibited by certain models where increased computation degrades performance.

Code Generation and Agentic Tasks. Figure 2 illustrates the comparative analysis of ARISE and Scaling Metric on code generation and agentic tasks across the Qwen3 model family. We observe a consistent positive correlation between model parameter size and both metrics, validating that larger models generally exhibit superior test-time scaling capabilities. Interestingly, code generation tasks consistently achieve higher ARISE scores compared to agentic tasks across all model sizes, suggesting that programming problems benefit more

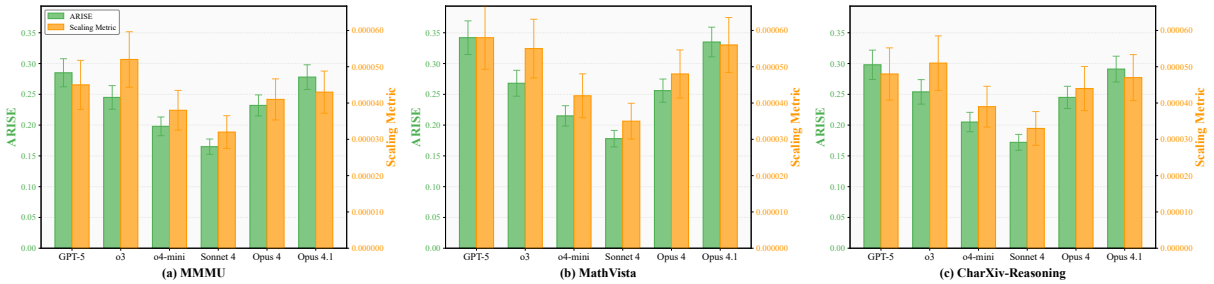

Figure 3: Comparison of ARISE and Scaling Metric across state-of-the-art reasoning models on multimodal reasoning tasks. The evaluation encompasses three challenging benchmarks: (a) MMMU, (b) MathVista, and (c) CharXiv-Reasoning. ARISE values are shown on the left y-axis (green bars) while Scaling Metric values are shown on the right y-axis (orange bars).

substantially from test-time scaling than tool-use.

Multimodal Reasoning. Figure 3 presents our analysis on multimodal reasoning tasks. GPT-5 and Claude Opus 4.1 achieve the highest performance across all three benchmarks, with GPT-5 demonstrating a marginal advantage on CharXiv-Reasoning, particularly for complex scientific chart interpretation. Compared to the Scaling Metric, ARISE provides significantly better discriminative power. The performance gap between weaker models (e.g., Sonnet 4) and stronger models (e.g., Opus 4.1) becomes more pronounced under ARISE evaluation, with differences exceeding 75.2% versus 45.6% in Scaling Metrics.

4.3 Analysis

Penalization of Performance Degradation. Table 1 reveals a critical distinction between ARISE and Scaling Metric in handling negative scaling behaviors. While both metrics generally provide concordant evaluations, ARISE uniquely captures performance degradation through negative scores. For instance, GPT-OSS-20B achieves -0.403 on AIME and DeepSeek-R1 scores -0.049 on GPQA Diamond, indicating that many initially correct samples become incorrect with increased computation, a phenomenon consistent with recent findings by Zeng et al. (2025). Scaling Metric remains insensitive to this crucial behavior, assigning only positive values regardless of degradation, failing to effectively penalize computational waste. We provide a detailed analysis of sample accuracy transitions before and after scaling in Appendix F.

Evaluation Stability. The error bars in Figure 2 demonstrate ARISE's superior stability compared to traditional metrics. When model performances are obvious (e.g., 32B, 30B, and 235B variants), ARISE exhibits notably smaller variance. To quantify this stability independent of scale differences, we compute the coefficient of variation (CV) across five independent runs. ARISE achieves an average CV of 0.14, substantially lower than Scaling Metric's 0.28. This pattern extends to multimodal tasks (Figure 3), where ARISE's average CV of 0.08 outperforms Scaling Metric's 0.15. The enhanced stability stems from our adaptive sampling mechanism, which dynamically allocates computational budget based on observed variance patterns.

Cross-Dataset Consistency. Table 1 demonstrates ARISE's remarkable consistency across diverse evaluation domains. For instance, Claude Opus 4.1 maintains ARISE scores within a narrow range (average deviation < 0.015) across mathematical and scientific reasoning tasks, despite their distinct problem characteristics. To quantify cross-dataset consistency, we compute the coefficient of variation across different benchmarks for each model. ARISE achieves an average inter-dataset CV of 0.194, compared to 0.249 for traditional Scaling Metric. This consistency indicates that ARISE captures fundamental scaling properties that transcend specific task domains, providing a more robust assessment of model capabilities.

Model Evolution Tracking. Our experiments reveal clear progression patterns across model generations, validating ARISE's sensitivity to architectural improvements. The OpenAI o-series progresses from o1 (average ARISE ≈ 0.134) to o3 (≈ 0.239), representing a 78% improvement. More dramatically, the Anthropic Claude series shows substantial gains from Claude Sonnet 4 (≈ 0.079) through Claude Opus 4 (≈ 0.268) to Claude Opus 4.1 (≈ 0.465), achieving nearly 6× improvement. These patterns, consistently observed in both text-based (Table 1) and multimodal evaluations (Figure 3), demonstrate that recent model development

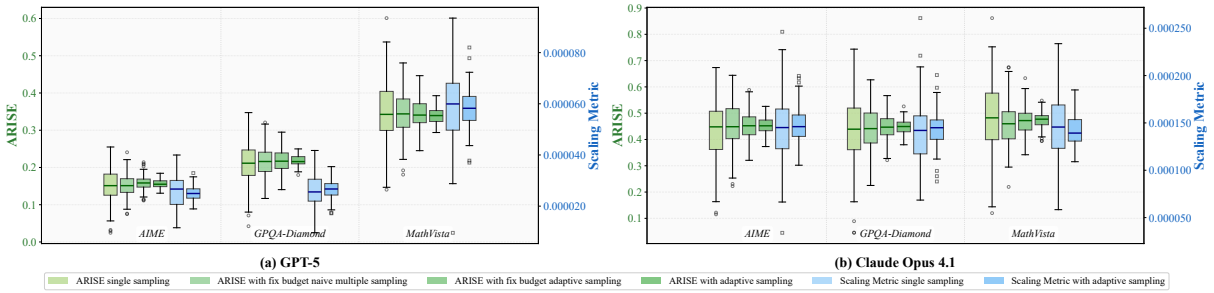

Figure 4: Stability analysis of adaptive sampling on (a) GPT-5 and (b) Claude Opus 4.1. Box plots compare variance across six sampling strategies, demonstrating that adaptive sampling achieves superior stability.

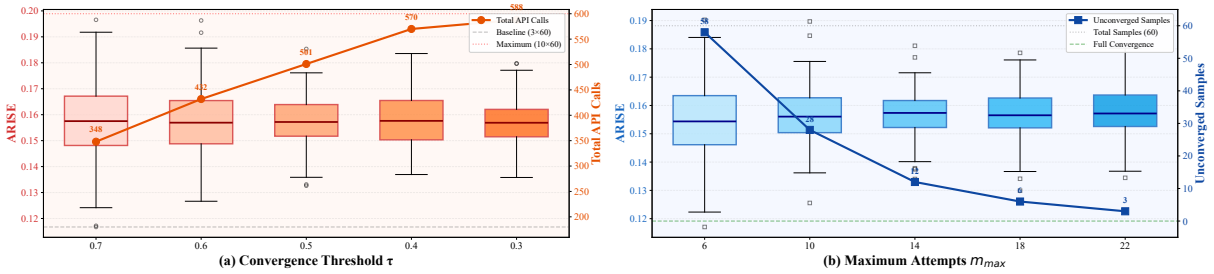

Figure 5: Hyperparameter analysis of adaptive sampling using GPT-5 on AIME. (a) Threshold τ: box plots show ARISE (left axis) while the line plot indicates total API calls (right axis). (b) Maximum attempts m_{max}: box plots display ARISE (left axis) while the line plot tracks unconverged samples failing to meet threshold τ (right axis).

has successfully enhanced test-time scaling capabilities and validate ARISE's effectiveness in tracking these advancements.

4.4 Adaptive Sampling Effectiveness

Impact of Adaptive Sampling on Evaluation Stability. Figure 4 presents a comprehensive analysis of how adaptive sampling affects evaluation stability across different datasets and models. The results demonstrate that adaptive sampling substantially enhances stability for both metrics across all evaluated benchmarks. Notably, while ARISE exhibits comparable or even higher variance than Scaling Metric under single sampling, particularly on GPQA-Diamond, the introduction of adaptive sampling reverses this relationship. With full adaptive sampling enabled, ARISE achieves remarkable stability improvements with variance reduction of 76.1%, significantly outperforming Scaling Metric's 54.2% reduction. This superior improvement stems from ARISE's sample-level awareness, which enables more targeted allocation of computational resources to high-variance instances.

Fixed Budget Analysis. To evaluate the efficiency of adaptive sampling under resource constraints, we conduct experiments with fixed computational budgets, comparing adaptive sampling against naive multiple sampling that uniformly distributes resources across all instances. As illustrated in Figure 4, adaptive sampling consistently outperforms naive approaches at all configurations. Under fixed budget constraints, naive multiple sampling achieves a modest 31.4% variance reduction compared to single sampling, while adaptive sampling delivers a substantial 57.5% reduction, nearly doubling the effectiveness despite identical costs. This efficiency gain is particularly pronounced for challenging instances where performance exhibits high variability. The adaptive mechanism identifies these unstable cases and allocates additional samples accordingly, whereas naive sampling wastes resources on already-stable predictions.

4.5 Hyperparameter Analysis

Convergence Threshold. Figure 5(a) examines the trade-off between evaluation stability and computational cost for convergence threshold τ. As τ decreases, variance monotonically decreases while sampling counts increase substantially. However, diminishing returns become evident beyond $\tau = 0.5$, indicating stricter thresholds provide minimal stability gains at significantly higher computational cost. This establishes $\tau = 0.5$ as the optimal operating point, balancing effective variance reduction with reasonable resource consumption.

Maximum Sampling Attempts. Figure 5(b) analyzes the effect of maximum attempts m_{max} on convergence behavior. While increasing m_{max} pro-

gressively reduces variance, the improvement rate declines after $m_{max} = 10$, as evidenced by the rapidly decreasing number of unconverged samples. Extended sampling beyond this threshold yields marginal benefits while potentially doubling computation for difficult instances. These findings confirm that $m_{max} = 10$ optimally balances stability and efficiency.

Diagnostic Sensitivity and Robustness. Beyond static alignment with human judgment, we further evaluate whether ARISE achieves an ideal diagnostic profile through a controlled perturbation study. Starting from the original scaling trajectories of GPT-5 and Claude Opus 4.1 on AIME, we introduce two types of perturbations. The first injects negative scaling by artificially flipping $k\%$ of correctly solved samples to incorrect at the next scaling level, with k varying from 5 to 25. The second injects Gaussian noise into token counts without altering accuracy, with the noise magnitude σ ranging from 0 to 0.3μ. A well-designed metric should be highly responsive to the first type, which represents genuine degradation in scaling quality, and robust to the second type, which represents irrelevant measurement noise.

As shown in Figure 6, ARISE exhibits strong sensitivity to negative scaling injection: a 10% injection rate leads to a 27.1% decrease in ARISE, whereas Scaling Metric changes by only 3.5%, yielding a 7.7× amplification factor. Meanwhile, ARISE demonstrates superior robustness to token noise, maintaining a coefficient of variation below 0.13 even at 30% noise levels, compared to Scaling Metric's 0.31. This 2.4× robustness advantage confirms that ARISE does not trade sensitivity for instability. Together, these results establish that ARISE achieves the ideal diagnostic profile of high responsiveness to genuine scaling quality differences coupled with strong resilience to measurement noise.

Diagnostic Sensitivity and Robustness. We further evaluate whether ARISE achieves an ideal diagnostic profile through a controlled perturbation study. Starting from the original scaling trajectories of GPT-5 and Claude Opus 4.1 on AIME, we introduce two types of perturbations. The first injects negative scaling by artificially flipping $k\%$ of correctly solved samples to incorrect at the next scaling level, with k varying from 5 to 25. The second injects Gaussian noise into token counts without altering accuracy, with the noise magnitude σ ranging from 0 to 0.3μ. A well-designed metric should be highly responsive to the first type, which represents genuine degradation in scaling quality, and robust to the second type, which represents irrelevant measurement noise.

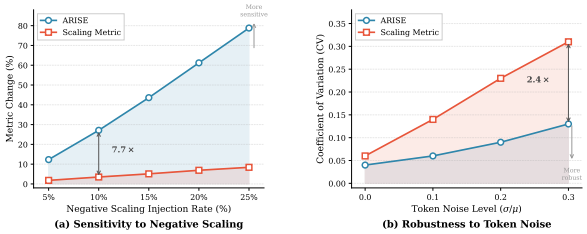

Figure 6: Diagnostic sensitivity-robustness profile of ARISE versus Scaling Metric via controlled perturbation on AIME. (a): sensitivity to negative scaling injection by flipping $k\%$ of correct samples to incorrect, where ARISE exhibits 7.7× higher responsiveness. (b): robustness to Gaussian token noise injection, where ARISE maintains 2.4× lower coefficient of variation.

As shown in Figure 6, ARISE exhibits strong sensitivity to negative scaling injection: a 10% injection rate leads to a 27.1% decrease in ARISE, whereas Scaling Metric changes by only 3.5%, yielding a 7.7× amplification factor. Meanwhile, ARISE demonstrates superior robustness to token noise, maintaining a coefficient of variation below 0.13 even at 30% noise levels, compared to Scaling Metric's 0.31. This 2.4× robustness advantage confirms that ARISE does not trade sensitivity for instability. Together, these results establish that ARISE achieves the ideal diagnostic profile of high responsiveness to genuine scaling quality differences coupled with strong resilience to measurement noise.

5 Conclusion

We propose ARISE, a novel metric for evaluating test-time scaling capabilities that addresses fundamental limitations of existing approaches through sample-level awareness and adaptive sampling mechanisms. By tracking individual sample trajectories across scaling iterations and appropriately penalizing performance degradation, ARISE provides principled assessment of scaling effectiveness while the dynamic sampling strategy ensures statistically reliable measurements under variance. Experiments across diverse reasoning tasks demonstrate that ARISE delivers consistent, discriminative and stable evaluations, establishing it as an principled framework for comparing test-time scaling capabilities across large reasoning models.

Limitations and Broader Impacts

Generalization to Diverse Scenarios. While we conducted extensive evaluation across mathematical, scientific, coding, agentic, and multimodal tasks, the test-time scaling capabilities of reasoning models in more scenarios remain unexplored. For instance, more diverse agentic environments beyond the mock, airline, retail, and telecom scenarios in τ^2-bench (Barres et al., 2025) is necessary to explore. Although expanding evaluation scope incurs additional computational costs, ARISE demonstrates strong cross-task consistency, suggesting that incorporating more evaluation domains will progressively approximate models' true test-time scaling capabilities.

Cross-Linguistic Evaluation. Our evaluation is currently limited to English-language datasets due to resource constraints. Cross-linguistic analysis is essential as AI systems should equitably serve users across different languages rather than privileging English speakers. Previous research has identified significant performance disparities in large language models when solving identical problems across different languages (Shafayat et al., 2024; Xuan et al., 2025). Future work should evaluate test-time scaling behaviors of large reasoning models across diverse languages when appropriate evaluation datasets become available, ensuring that scaling benefits are universally accessible.

Ethics Statement

Data Usage Compliance. Throughout our experiments, we strictly adhere to all applicable data usage regulations and licensing requirements. Table 2 provides comprehensive details for each dataset, including domain, answer type, sample size, and corresponding license information. All datasets utilized in this work are in English, and we have carefully verified that our usage complies with the specific licensing terms of each dataset.

Model Usage Compliance. We maintain strict adherence to all model usage regulations and terms of service. Table 3 details the specifications of each model, including parameter counts, release dates, and releasing organizations. For proprietary models, we strictly comply with the terms of service established by their respective organizations and exclusively utilize official APIs. For open-source models, we ensure full compliance with their corresponding licenses and usage guidelines.

Safety and Content Moderation. We implement rigorous safeguards to ensure that all model outputs remain safe and appropriate, containing no personal information or offensive content. The datasets employed in our evaluation have undergone thorough community validation and review, and we have identified no instances of unsafe or inappropriate content throughout our experimental process.

Use of AI Assistants. During the development of this work, we utilized Cursor for code composition and development. We ensure that all usage of AI tools strictly complies with submission guidelines and ethical standards established by the ACL community.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. U24B20181 and 62525602).

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*.

Anthropic. 2025a. Claude opus 4.1.

Anthropic. 2025b. Introducing claude 4.

Vidhisha Balachandran, Jingya Chen, Lingjiao Chen, Shivam Garg, Neel Joshi, Yash Lara, John Langford, Besmira Nushi, Vibhav Vineet, Yue Wu, et al. 2025. Inference-time scaling for complex tasks: Where we stand and what lies ahead. *arXiv preprint arXiv:2504.00294*.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. Matharena: Evaluating llms on uncontaminated math competitions.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2-bench: Evaluating conversational agents in a dual-control environment. *Preprint*, arXiv:2506.07982.

Aman Bhargava, Cameron Witkowski, Shi-Zhuo Looi, and Matt Thomson. 2023. What's the magic word? a control theory of llm prompting. *arXiv preprint arXiv:2310.04444*.

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*.

Souradip Chakraborty, Mohammadreza Pourreza, Ruoxi Sun, Yiwen Song, Nino Scherrer, Furong Huang, Amrit Singh Bedi, Ahmad Beirami, Jindong Gu, Hamid Palangi, and Tomas Pfister. 2025. On the role of feedback in test-time scaling of agentic ai workflows. *Preprint*, arXiv:2504.01931.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

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

Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024b. Simple and provable scaling laws for the test-time compute of large language models. *arXiv preprint arXiv:2411.19477*.

Cheng-Han Chiang and Hung-yi Lee. 2024. Over-reasoning and redundant calculation of large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 161–169, St. Julian's, Malta. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Shihan Dou, Ming Zhang, Chenhao Huang, Jiayi Chen, Feng Chen, Shichun Liu, Yan Liu, Chenxiao Liu, CHENG ZHONG, Zongzhang Zhang, et al. Evalearn: Quantifying the learning capability and efficiency of llms via sequential problem solving. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, et al. 2026. Clbench: A benchmark for context learning. *arXiv preprint arXiv:2602.03587*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *Preprint*, arXiv:2503.24290.

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

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? In *ICLR*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Dacheng Li, Shiyi Cao, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E Gonzalez, and Ion Stoica. 2025. S*: Test time scaling for code generation. *arXiv preprint arXiv:2502.14382*.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

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*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025a. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025b. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. Notion Blog.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

OpenAI. 2025a. Gpt-5 is here.

OpenAI. 2025b. Introducing openai o3 and o4-mini.

OpenAI. 2025c. Introducing openai o3 and o4-mini.

OpenAI. 2025d. Openai o3-mini.

Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*.

Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. 2024. mhumaneval–a multilingual benchmark to evaluate large language models for code generation. *arXiv preprint arXiv:2410.15037*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-FAct: Assessing factuality of multilingual LLMs using FActscore. In *First Conference on Language Modeling*.

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

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. 2025. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Computing Surveys*, 57(11):1–43.

Qiushi Sun, Zhirui Chen, Fangzhi Xu, Kanzhi Cheng, Chang Ma, Zhangyue Yin, Jianing Wang, Chengcheng Han, Renyu Zhu, Shuai Yuan, et al. 2024. A survey of neural code intelligence: Paradigms, advances and beyond. *arXiv preprint arXiv:2403.14734*.

AoPSWiki Team. 2025. Aime problems and solutions.

Fengwei Teng, Zhaoyang Yu, Quan Shi, Jiayi Zhang, Chenglin Wu, and Yuyu Luo. 2025. Atom of thoughts for markov llm test-time scaling. *arXiv preprint arXiv:2502.12018*.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*.

Jian Wang, Boyan Zhu, Chak Tou Leong, Yongqi Li, and Wenjie Li. 2025a. Scaling over scaling: Exploring test-time scaling pareto in large reasoning models. *arXiv preprint arXiv:2505.20522*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024a. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

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

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Guojun Wu. 2025. It's not that simple. an analysis of simple test-time scaling. *arXiv preprint arXiv:2507.14419*.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. In *The Thirteenth International Conference on Learning Representations*. OpenReview.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *Preprint*, arXiv:2503.10497.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhangyue Yin, Qiushi Sun, Zhiyuan Zeng, Qinyuan Cheng, Xipeng Qiu, and Xuanjing Huang. 2025. Dynamic and generalizable process reward modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4203–4233, Vienna, Austria. Association for Computational Linguistics.

Zhaojian Yu, Yinghao Wu, Yilun Zhao, Arman Cohan, and Xiao-Ping Zhang. 2025. Z1: Efficient test-time scaling with code. *arXiv preprint arXiv:2504.00810*.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. 2024. Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective. *arXiv preprint arXiv:2412.14135*.

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? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4651–4665, Vienna, Austria. Association for Computational Linguistics.

Ming Zhang, Kexin Tan, Yueyuan Huang, Yujiong Shen, Chunchun Ma, Li Ju, Xinran Zhang, Yuhui Wang, Wenqing Jing, Jingyi Deng, et al. 2026. Opennovelty: An llm-powered agentic system for verifiable scholarly novelty assessment. *arXiv preprint arXiv:2601.01576*.

Ming Zhang, Yue Zhang, Shichun Liu, Haipeng Yuan, Junzhe Wang, Yurui Dong, Jingyi Deng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Llmeval-2.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, et al. 2025. A survey on test-time scaling in large language models: What, how, where, and how well? *arXiv preprint arXiv:2503.24235*.

Yue Zhang, Ming Zhang, Haipeng Yuan, Shichun Liu, Yongyao Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Llmeval: A preliminary study on how to evaluate large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19615–19622.

King Zhu, Hanhao Li, Siwei Wu, Tianshun Xing, Dehua Ma, Xiangru Tang, Minghao Liu, Jian Yang, Jiaheng Liu, Yuchen Eleanor Jiang, et al. 2025. Scaling test-time compute for llm agents. *arXiv preprint arXiv:2506.12928*.

A Boundedness Analysis

We provide formal boundedness proofs for both the traditional Scaling Metric and our proposed ARISE metric, demonstrating the fundamental differences in their mathematical properties.

A.1 Scaling Metric Bounds

The Scaling Metric (Muennighoff et al., 2025) is defined as:

$$\text{Scaling} = \frac{1}{\binom{|\mathcal{P}|}{2}} \sum_{\substack{p_1,p_2 \in \mathcal{P} \\ \mathcal{T}(p_2) > \mathcal{T}(p_1)}} \frac{\mathcal{A}(p_2) - \mathcal{A}(p_1)}{\mathcal{T}(p_2) - \mathcal{T}(p_1)} \tag{18}$$

Theorem 1 (Scaling Metric Bounds) *The Scaling Metric is bounded: Scaling* $\in \left[\frac{-1}{\delta_{\min}}, \frac{1}{\delta_{\min}}\right]$ *where* $\delta_{\min} = \min_{p_1,p_2}(\mathcal{T}(p_2) - \mathcal{T}(p_1))$.

Proof 1 *For any pair* (p_1, p_2) *with* $\mathcal{T}(p_2) > \mathcal{T}(p_1)$*:*

Since $\mathcal{A} \in [0, 1]$*, the numerator satisfies:*

$$-1 \leq \mathcal{A}(p_2) - \mathcal{A}(p_1) \leq 1 \tag{19}$$

For the denominator, by definition:

$$\mathcal{T}(p_2) - \mathcal{T}(p_1) \geq \delta_{\min} > 0 \tag{20}$$

Combining Equations 19 and 20:

$$\frac{-1}{\delta_{\min}} \leq \frac{\mathcal{A}(p_2) - \mathcal{A}(p_1)}{\mathcal{T}(p_2) - \mathcal{T}(p_1)} \leq \frac{1}{\delta_{\min}} \tag{21}$$

Since the Scaling Metric is an average over all pairs, it inherits these bounds. In practice, with normalized token differences where $\delta_{\min} \to 1^+$*, we obtain Scaling Metric* $\in (-1, 1)$.

A.2 ARISE Bounds

For ARISE, we analyze bounds considering all possible accuracy transition patterns.

Theorem 2 (Sample-Level ARISE Bounds) *For a single sample* i *with* m *scaling iterations,* $\text{ARISE}_i \in (-\infty, 1)$.

Proof 2 *From Equation 5,* ARISE_i *is the sum of contributions from all transitions. Let* $\mathcal{I} = \{j : a_i^{(j)} = 1, a_i^{(j-1)} = 0\}$ *denote improvement transitions and* $\mathcal{D} = \{j : a_i^{(j)} = 0, a_i^{(j-1)} = 1\}$ *denote degradation transitions.*

$$ARISE_i = \sum_{j \in \mathcal{I}} \frac{t_i^{(j-1)}}{t_i^{(j)}} - \sum_{j \in \mathcal{D}} \frac{t_i^{(j)}}{t_i^{(j-1)}} \tag{22}$$

Upper bound: *The maximum occurs when there are only improvements and no degradations. The optimal accuracy sequence with a single transition at position j^*:*

$$ARISE_{i,\max} = \frac{t_i^{(j^*-1)}}{t_i^{(j^*)}} < 1 \tag{23}$$

As $t_i^{(j^*)} \to t_i^{(j^*-1)}$, $ARISE_{i,\max} \to 1^-$. Thus, the supremum is 1.

Lower bound: *The minimum occurs with only degradations and no improvements. The worst accuracy sequence with a single transition at position j^*:*

$$ARISE_{i,\min} = -\frac{t_i^{(j^*)}}{t_i^{(j^*-1)}} < -1 \tag{24}$$

Since $t_i^{(j^*)}$ can be arbitrarily larger than $t_i^{(j^*-1)}$, $ARISE_{i,\min} \to -\infty$, yielding intermediate values within $(-\infty, 1)$.

Theorem 3 (ARISE Bounds) *The aggregate ARISE metric satisfies $ARISE \in (-\infty, 1)$.*

Proof 3 *From Equation 4, ARISE is the arithmetic mean of sample scores:*

$$ARISE = \frac{1}{n} \sum_{i=1}^{n} ARISE_i \tag{25}$$

By Theorem 2, each $ARISE_i \in (-\infty, 1)$.

Upper bound: *When all samples achieve the optimal pattern :*

$$ARISE_{\max} = \lim_{t_i^{(j^*)} \to t_i^{(j^*-1)}} \frac{1}{n} \sum_{i=1}^{n} \frac{t_i^{(j^*-1)}}{t_i^{(j^*)}} < 1 \tag{26}$$

Lower bound: *When all samples exhibit the worst pattern :*

$$ARISE_{\min} = \lim_{r \to \infty} \frac{1}{n} \sum_{i=1}^{n} (-r) = -\infty \tag{27}$$

where $r = \frac{t_i^{(j^*)}}{t_i^{(j^*-1)}}$ is the token ratio. Thus, the range of the ARISE is $(-\infty, 1)$.

Algorithm 1 ARISE Computation with Adaptive Sampling

Require: Dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$, model $\mathcal{M}$, scaling iterations J, budget B
Ensure: ARISE score

```
Initialize ARISE ← 0
for i = 1 to n do                                   ▷ For each sample
    ARISE_i ← 0
    for j = 1 to J do                               ▷ Scaling iterations
        // Adaptive sampling strategy
        k*, a_i^(j), t_i^(j) ← AdaptiveSample(M, x_i, j)
        // Compute contribution
        Δa ← a_i^(j) − a_i^(j−1)
        if Δa ≠ 0 then
            W ← (t_i^(j−1) / t_i^(j))^sign(Δa)
            ARISE_i ← ARISE_i + Δa · W
        end if
    end for
    ARISE ← ARISE + ARISE_i
end for
return ARISE/n
```

B Algorithm

We present the complete algorithmic procedure for computing ARISE with adaptive sampling. Algorithm 1 outlines the main computation flow, while Algorithm 2 details the dynamic sampling strategy.

B.1 Algorithmic Analysis

Computational Complexity. The time complexity of ARISE is $\mathcal{O}(n \cdot J \cdot \bar{k} \cdot C_{\text{eval}})$, where n is the dataset size, J is the number of scaling iterations, $\bar{k}$ is the average sample count per configuration, and C_{eval} is the cost of a single model evaluation. The adaptive termination in Algorithm 2 ensures $\bar{k} \ll m_{\max}$ for stable configurations, significantly reducing computation compared to fixed sampling.

Early Termination. Algorithm 2 implements dynamic early stopping based on the coefficient of variation (CV) from accuracy and token statistics (line 14). When CV $< \tau$, sampling terminates due to sufficient statistical stability. Low-variance configurations converge with $k \approx m_{\min}$ samples, while high-variance cases automatically expand to $k \to m_{\max}$.

Adaptive Resource Allocation. The sample count k^* varies per configuration based on ob-

Algorithm 2 Adaptive Sampling Strategy

Require: Model $\mathcal{M}$, sample x_i, iteration j, threshold τ, max trials $m_{\max}$
Ensure: Sample count k^*, mean accuracy $a_i^{(j)}$, mean tokens $t_i^{(j)}$

// Initial probing phase
$k \leftarrow m_{\min}$ ▷ Minimum trials
for $\ell = 1$ to $m_{\min}$ **do**
  $a_{i,\ell}^{(j)} \leftarrow \textsc{Evaluate}(\mathcal{M}, x_i, t^{(j)})$
  $t_{i,\ell}^{(j)} \leftarrow \textsc{CountTokens}(\mathcal{M}, x_i)$
end for
// Compute initial statistics
$\mu_a \leftarrow \frac{1}{k}\sum_{\ell=1}^{k} a_{i,\ell}^{(j)}$
$\sigma_a \leftarrow \sqrt{\frac{1}{k}\sum_{\ell=1}^{k}(a_{i,\ell}^{(j)} - \mu_a)^2}$
$\mu_t \leftarrow \frac{1}{k}\sum_{\ell=1}^{k} t_{i,\ell}^{(j)}$
$\sigma_t \leftarrow \sqrt{\frac{1}{k}\sum_{\ell=1}^{k}(t_{i,\ell}^{(j)} - \mu_t)^2}$
// Compute coefficient of variation
$\text{CV} \leftarrow \frac{\sigma_a}{\mu_a+\epsilon} + \frac{\sigma_t}{\mu_t+\epsilon}$
while $\text{CV} \geq \tau$ **and** $k \leq m_{\max}$ **do**
  $k \leftarrow k + 1$
  $a_{i,k}^{(j)} \leftarrow \textsc{Evaluate}(\mathcal{M}, x_i, t^{(j)})$
  $t_{i,k}^{(j)} \leftarrow \textsc{CountTokens}(\mathcal{M}, x_i)$
  Update $\mu_a, \sigma_a, \mu_t, \sigma_t, \text{CV}$
end while
$k^* \leftarrow k$ ▷ Final sample count
$a_i^{(j)} \leftarrow \frac{1}{k^*}\sum_{\ell=1}^{k^*} a_{i,\ell}^{(j)}$
$t_i^{(j)} \leftarrow \frac{1}{k^*}\sum_{\ell=1}^{k^*} t_{i,\ell}^{(j)}$
return $k^*, a_i^{(j)}, t_i^{(j)}$

served variance, enabling natural budget redistribution. Stable configurations converge with minimal samples, while stochastic configurations adaptively require more trials, concentrating resources where uncertainty is highest.

C Experiment Details

C.1 Inference Configuration

For open-source models, we deploy them using vLLM (Kwon et al., 2023) for efficient inference. Following official recommendations, we configure the generation hyperparameters as follows: Temperature = 0.6, Top-P = 0.95, Top-K = 20, and Min-P = 0. These settings balance between generation diversity and output quality, enabling controlled yet flexible reasoning processes during test-time scaling evaluation.

C.2 Test-Time Scaling Configuration

For models supporting variable test-time computation, we evaluate across multiple scaling levels:

- **Effort-based scaling** (OpenAI o-series, gpt-oss models): We test `low`, `medium`, and `high` reasoning effort levels, allowing models to allocate increasing computational resources to complex problems.
- **Mode-based scaling** (Claude 4 series, Qwen3 series and DeepSeek models): We compare `think` mode (with explicit reasoning chains) against `no-think` mode (direct response generation). For DeepSeek models, we utilize `reasoner` (`think`) and standard `chat` (`no-think`) modes for comparative analysis, with explicit prompts employed to control reasoning length for DeepSeek-R1.

All experiments are conducted with a maximum token limit of 32,768 for reasoning traces and 2,048 for queries, ensuring sufficient space for complex multi-step reasoning while maintaining computational feasibility. For consistency, we perform five independent runs for each configuration and report averaged results with standard deviations.

D Dataset Details

In our experiments, we selected eleven datasets encompassing a comprehensive range of task types that require sophisticated reasoning capabilities and test-time scaling behaviors. These datasets were specifically chosen to evaluate different aspects of reasoning model performance across mathematical, scientific, coding, agentic, and multimodal domains. Detailed statistics for these datasets are provided in Table 2.

D.1 Mathematical Reasoning Datasets

- **AIME** (Team, 2025) (American Invitational Mathematics Examination) consists of 60 challenging mathematics problems selected from the 2024 and 2025 competitions (30 problems each year). Each problem requires an integer answer between 0 and 999, enabling precise evaluation of mathematical reasoning using Pass@1 as the metric. The problems span advanced topics including algebra, geometry, number theory, combinatorics, and probability, requiring multi-step reasoning and creative problem-solving approaches.

Dataset	Domain	Answer Format	# Samples	License
Mathematical Reasoning				
AIME (Team, 2025)	Competition Math	Integer (0-999)	60	CC BY-NC-SA 4.0
HMMT (Balunović et al., 2025)	Competition Math	Free-form	60	CC BY-NC-SA 4.0
Scientific Reasoning				
GPQA-Diamond (Rein et al., 2024)	Graduate Science	Multi-Choice (4)	198	CC BY 4.0
MMLU-Pro (Wang et al., 2024a)	Professional Knowledge	Multi-Choice (10)	12,032	MIT
Code Generation				
SWE-bench Verified (Jimenez et al., 2024)	Software Engineering	Code Patch	500	CC0 1.0
LiveCodeBench (Jain et al., 2025)	Code Generation	Full Code	121	CC
Agentic Tasks				
τ^2-Bench (Barres et al., 2025)	Conversational Agents	Tool Calls	279	MIT
BFCL-v3 (Patil et al., 2025)	Function Calling	JSON	4,441	Apache 2.0
Multimodal Reasoning				
MMMU (Yue et al., 2024)	Multimodal Understanding	Multi-Choice	11,550	Apache 2.0
MathVista (Lu et al., 2024)	Mathematical Visual	Mixed	6,141	CC BY-SA 4.0
CharXiv-Reasoning (Wang et al., 2024b)	Chart Understanding	Free-form	2,323	CC BY-SA 4.0

Table 2: Statistics of evaluation datasets used in our ARISE experiments. Datasets span mathematical reasoning, scientific understanding, code generation, agentic tasks, and multimodal reasoning domains.

- **HMMT** (Balunović et al., 2025) (Harvard-MIT Mathematics Tournament) comprises 60 problems from the February 2024 and February 2025 competitions. Unlike AIME's integer-only format, HMMT accepts free-form mathematical expressions including fractions, algebraic expressions, and LaTeX-formatted answers, evaluated using Pass@1 metric. The dataset is categorized by problem type (combinatorics, algebra, geometry, and number theory).

D.2 Scientific Reasoning Datasets

- **GPQA-Diamond** (Rein et al., 2024) (Graduate-Level Google-Proof Q&A) contains 198 meticulously curated graduate-level science questions designed to be "Google-proof"—questions that cannot be easily answered through simple web searches. Each multiple-choice question has 4 options and was authored by PhD holders and validated by domain experts who achieved 81% accuracy, while non-experts with internet access reached only 22% accuracy after 30+ minutes of searching. The questions span biology, physics, and chemistry at the graduate level, with performance measured by accuracy.

- **MMLU-Pro** (Wang et al., 2024a) represents a substantial enhancement over the original MMLU benchmark, containing 12,032 questions with 10 answer choices (A-J) instead of the traditional 4, reducing random guess probability from 25% to 10%. The dataset covers 14 domains including STEM fields, humanities, social sciences, business, health, and law.

D.3 Code Generation Datasets

- **SWE-bench Verified** (Jimenez et al., 2024) provides 500 human-validated GitHub issues from 12 popular Python repositories, requiring models to generate code patches in git diff format to resolve real-world software engineering problems. Each issue is categorized by difficulty (15-minute "easy" to 1+ hour "hard" tasks) and has been verified by human developers to ensure solvability and clear problem specifications. The benchmark tests practical software engineering skills including debugging, feature implementation, and code refactoring, with performance measured by resolve rate.

- **LiveCodeBench** (Jain et al., 2025) offers 121 coding problems continuously collected from competitive programming platforms (LeetCode, AtCoder, CodeForces) after major models' training cutoffs, ensuring contamination-free evaluation. Problems require complete, executable Python code generation and are tested using Pass@1. The dynamic nature of the dataset, with problems added monthly, makes it particularly valuable for evaluating true generalization capabilities.

D.4 Agentic Task Datasets

- τ^2**-Bench** (Barres et al., 2025) evaluates conversational agents through 279 multi-turn dialogues across retail, airline, and telecom domains. The benchmark uniquely tests dual-control scenarios where both the agent and simulated user can modify shared state through tool calls and API interactions. Success is measured through Pass^k (Pass-Hat-K) metric assessing task completion rates, policy adherence, and database state matching, providing comprehensive assessment of agent coordination and planning capabilities.

- **BFCL-v3** (Patil et al., 2025) (Berkeley Function Calling Leaderboard v3) contains 4,441 function calling scenarios testing single-turn, multi-turn, and multi-step interactions. Models must generate properly formatted JSON function calls, with evaluation using Abstract Syntax Tree (AST) substring matching for single-turn scenarios and combined state-based and response-based evaluation for multi-turn entries. The benchmark includes relevance detection tasks where models must determine when not to call functions, testing both precision and recall in tool use.

D.5 Multimodal Reasoning Datasets

- **MMMU** (Yue et al., 2024) (Massive Multi-discipline Multimodal Understanding) comprises 11,550 college-level questions requiring joint visual and textual reasoning across 30 subjects and 183 subfields. The dataset incorporates over 30 heterogeneous image types including charts, diagrams, maps, tables, chemical structures, and music sheets. Questions span six core disciplines (Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, Tech & Engineering), testing expert-level multimodal understanding with micro-averaged accuracy as the evaluation metric.

- **MathVista** (Lu et al., 2024) provides 6,141 mathematical reasoning problems in visual contexts, supporting both multiple-choice and free-form answers (integer, float, text, or list) evaluated by accuracy. The benchmark tests seven reasoning types: algebraic, arithmetic, geometric, logical, numeric common sense, scientific, and statistical reasoning. Problems are derived from 28 existing datasets plus three newly created sources, emphasizing mathematical reasoning grounded in visual information.

- **CharXiv-Reasoning** (Wang et al., 2024b) consists of 2,323 high-resolution scientific charts extracted from arXiv preprints, generating 11,615 questions (2,323 reasoning-focused). The dataset emphasizes complex chart understanding requiring synthesis across multiple visual elements, trend analysis, and comparative reasoning, with accuracy as the evaluation metric. Unlike simpler chart QA datasets, CharXiv features realistic scientific visualizations with complex legends, multiple subplots, and domain-specific annotations typical of academic publications.

E Model Details

We evaluate 22 state-of-the-art large reasoning models spanning four major organizations: OpenAI, Anthropic, Alibaba (Qwen), and DeepSeek. These models represent the current frontier of test-time scaling capabilities, ranging from compact 0.6B parameter models to massive 671B parameter mixture-of-experts architectures. Comprehensive specifications including parameter counts, release dates, context windows, and organizational affiliations are provided in Table 3. The evaluated models employ two primary test-time scaling approaches:

- **Effort-based scaling**: Models dynamically adjust computational resources during inference through configurable reasoning effort levels (`low`, `medium`, `high`). This approach, exclusive to OpenAI's o-series models (o1, o3, o3-mini, o4-mini) and open-weight gpt-oss variants (20B, 120B), allows users to explicitly control the depth of reasoning based on task complexity and latency requirements.

- **Mode-based scaling**: Models switch between distinct reasoning modes through specialized prompts. All non-OpenAI models in our evaluation, including Anthropic's Claude 4 series, Alibaba's Qwen3 family, and DeepSeek's V3.1, V3.1-Terminus and V3.2-Exp, utilize this approach. These models toggle between standard response generation (`no-think` or `chat` mode) and enhanced reasoning with

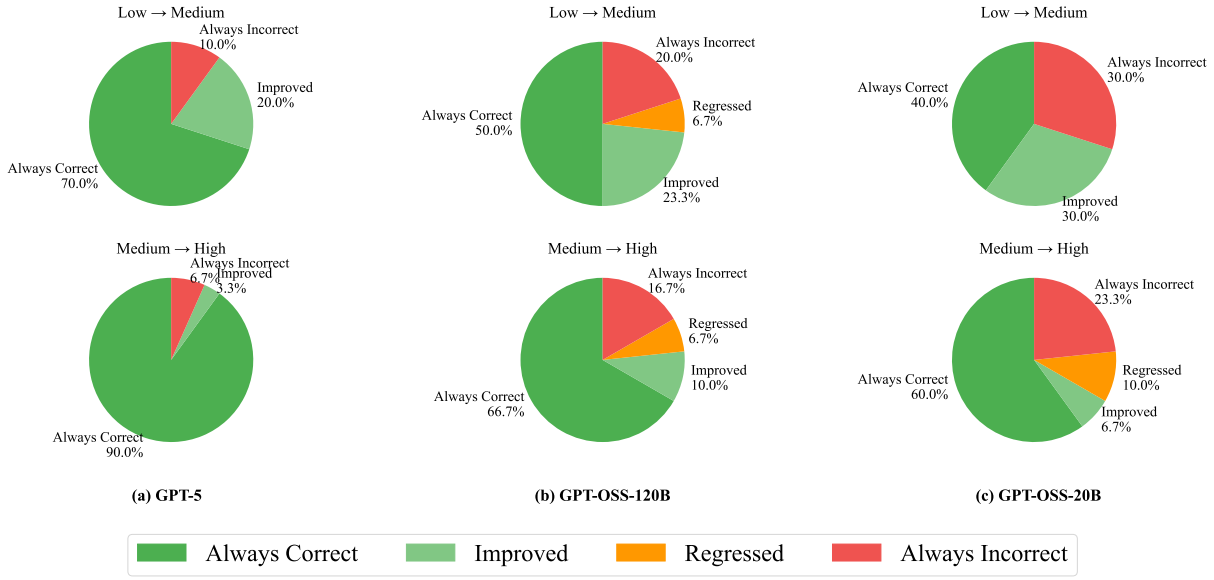

Figure 7: Sample-level accuracy transitions during test-time scaling on AIME dataset. We track individual sample accuracy changes across scaling iterations for GPT-5, GPT-OSS-120B, and GPT-OSS-20B. GPT-OSS models exhibit substantial performance degradation where many initially correct samples become incorrect.

explicit chain-of-thought traces (`think` or `reasoner` mode).

The diversity in model architectures, from dense transformers to mixture-of-experts, combined with varying test-time scaling mechanisms, provides a comprehensive landscape for evaluating the effectiveness of our proposed ARISE metric across different computational paradigms and reasoning strategies.

F Further Analysis

Effective Capture of Performance Degradation During Scaling. Figure 7 presents our analysis of sample-level accuracy transitions for GPT series models on the AIME dataset. Through tracking individual sample trajectories across scaling levels, we identify severe performance degradation in GPT-OSS-120B and GPT-OSS-20B. Specifically, GPT-OSS-120B exhibits 6.7% of samples transitioning from correct to incorrect when scaling from low to medium computational budget, with another 6.7% degrading during the medium-to-high transition. GPT-OSS-20B demonstrates even more pronounced instability, with 10.0% of samples experiencing degradation during medium-to-high scaling. In contrast, GPT-5 maintains consistent improvement throughout the scaling process without a single instance of correct samples becoming incorrect. These observations directly explain the significant negative ARISE scores observed for GPT-OSS-120B and GPT-OSS-20B in Table 1, demonstrating ARISE's unique ability to capture performance degradation that Scaling Metric completely overlooks.

Cross-Task Consistency. To further validate ARISE's consistency across different tasks, we conduct fine-grained analysis across MMLU-Pro's 14 discipline categories and MMMU's 6 core disciplines. As illustrated in Figure 8, both GPT-5 and Opus 4.1 exhibit remarkably consistent scaling patterns despite substantial task heterogeneity. Technical and reasoning-intensive disciplines consistently achieve the highest ARISE values, with Engineering leading for both models (GPT-5: 0.378, Opus 4.1: 0.568), followed by Computer Science and Mathematics. This pattern is corroborated by our MMMU analysis in Table 5, where Science consistently yields the highest ARISE scores while Humanities & Social Science exhibits lower scaling efficiency. Notably, the cross-discipline ranking remains stable across models, with GPT-5 (0.285) and Opus 4.1 (0.278) demonstrating clear advantages over other variants. The approximately 20% relative improvement from Opus 4 to Opus 4.1 highlights substantial progress in scaling efficiency across model generations. The cross-category coefficient of variation remains below 0.15 for both benchmarks, confirming ARISE's reliability across diverse knowledge domains.

Variance Analysis of Accuracy and Token Counts Figure 9 presents the analysis of the

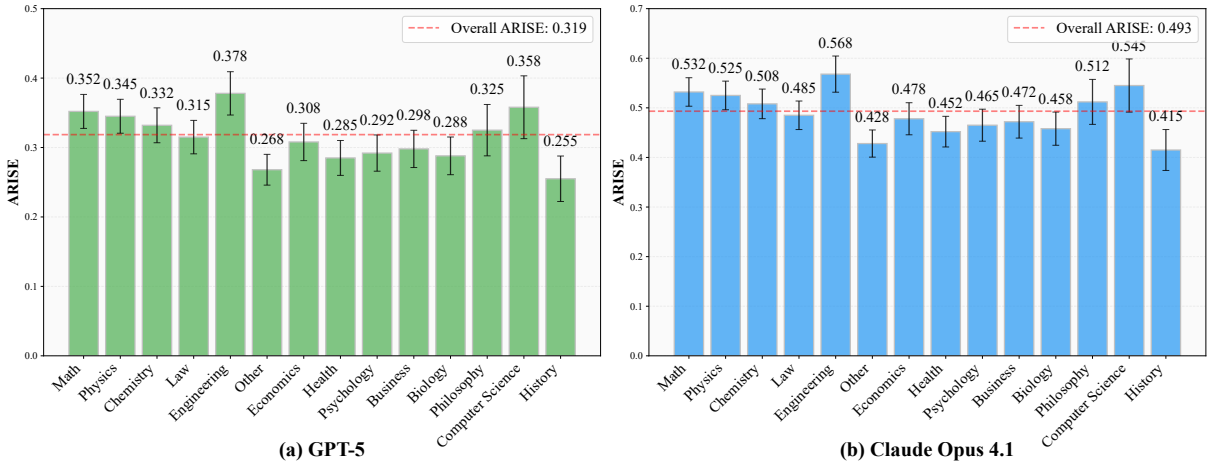

Figure 8: Category-level ARISE analysis across MMLU-Pro disciplines. Evaluation of (a) GPT-5 and (b) Claude Opus 4.1 on 14 distinct knowledge domains demonstrates consistent scaling patterns. Error bars represent standard errors. The red dashed line indicates the overall ARISE score across all categories.

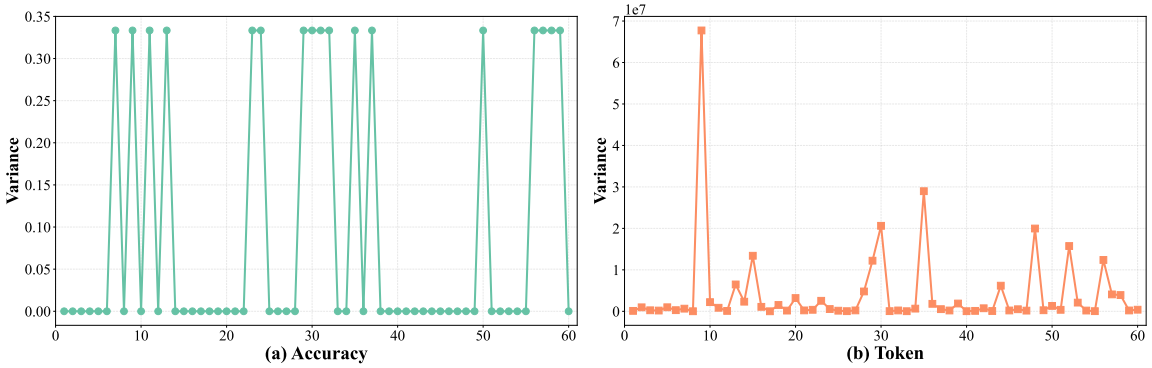

Figure 9: Variance analysis of sample accuracy and completion token counts during test-time scaling. Comparison of (a) accuracy variance and (b) token count variance for GPT-5 evaluated on the AIME dataset across different scaling levels. The completion token variance exhibits an order of magnitude of 10^7, significantly exceeding the accuracy variance.

variance in sample accuracy and completion token counts. On the AIME dataset, we observe notable fluctuations in accuracy across samples, which arise from the sampling of both correct and incorrect reasoning paths. While completion token variance exhibits substantial fluctuations in only a small subset of samples, its magnitude is several orders higher than that of accuracy variance. To address this scale disparity, we employ the coefficient of variation (CV) in Equation 12, which normalizes the variance by the mean. Our analysis on the AIME dataset yields $\text{CV}_a = 0.3175$ and $\text{CV}_t = 0.2854$ for GPT-5, demonstrating comparable magnitudes after normalization. This numerical proximity justifies our design choice in Equation 13, where we adopt a summation approach to compute the combined uncertainty metric that integrates both accuracy and token count variability.

Model	# Parameters	Release Date	Context Window	Scaling Type	Organization
OpenAI Models					
o1 (Jaech et al., 2024)	/	2024-09	128K	Effort	OpenAI
o3 (OpenAI, 2025b)	/	2025-04	200K	Effort	OpenAI
o3-mini (OpenAI, 2025d)	/	2025-01	200K	Effort	OpenAI
o4-mini (OpenAI, 2025c)	/	2025-04	200K	Effort	OpenAI
GPT-5 (OpenAI, 2025a)	/	2025-08	400K	Effort	OpenAI
gpt-oss-20B (Agarwal et al., 2025)	21B (3.6B active)	2025-08	128K	Effort	OpenAI
gpt-oss-120B (Agarwal et al., 2025)	117B (5.1B active)	2025-08	128K	Effort	OpenAI
Anthropic Models					
Claude Sonnet 4 (Anthropic, 2025b)	/	2025-05	200K	Mode	Anthropic
Claude Opus 4 (Anthropic, 2025b)	/	2025-05	200K	Mode	Anthropic
Claude Opus 4.1 (Anthropic, 2025a)	/	2025-08	200K	Mode	Anthropic
Qwen Models					
Qwen3-0.6B (Yang et al., 2025)	0.6B	2025-04	32K	Mode	Alibaba
Qwen3-1.7B (Yang et al., 2025)	1.7B	2025-04	32K	Mode	Alibaba
Qwen3-4B (Yang et al., 2025)	4B	2025-04	32K	Mode	Alibaba
Qwen3-8B (Yang et al., 2025)	8B	2025-04	128K	Mode	Alibaba
Qwen3-14B (Yang et al., 2025)	14B	2025-04	128K	Mode	Alibaba
Qwen3-32B (Yang et al., 2025)	32B	2025-04	128K	Mode	Alibaba
Qwen3-30B-A3B (Yang et al., 2025)	30B (3B active)	2025-04	128K	Mode	Alibaba
Qwen3-235B-A22B (Yang et al., 2025)	235B (22B active)	2025-04	256K	Mode	Alibaba
DeepSeek Models					
DeepSeek-R1 (DeepSeek-AI, 2025)	671B (37B active)	2025-01	128K	Mode	DeepSeek
DeepSeek-V3.1 (DeepSeek-AI, 2024)	671B (37B active)	2025-08	128K	Mode	DeepSeek
DeepSeek-V3.1-Terminus (DeepSeek-AI, 2024)	671B (37B active)	2025-09	128K	Mode	DeepSeek
DeepSeek-V3.2-Exp (DeepSeek-AI, 2024)	671B (37B active)	2025-09	128K	Mode	DeepSeek

Table 3: Statistics of large reasoning models. Models are categorized by their test-time scaling approach: effort-based (adjustable reasoning levels), and mode-based (thinking mode switching).

Model	**AIME**		**HMMT**		**GPQA Diamond**		**MMLU-Pro**	
	ARISE	SM	ARISE	SM	ARISE	SM	ARISE	SM
o1	0.134563	0.000060	0.127678	0.000018	0.122789	0.000038	0.150899	0.000050
o3	0.299253	0.000078	0.167327	0.000018	0.212377	0.000058	0.278925	0.000056
o3-mini	0.130586	0.000037	0.188770	0.000030	0.164882	0.000022	0.166285	0.000041
o4-mini	0.240192	0.000034	0.167306	0.000043	0.199441	0.000042	0.208032	0.000039
gpt-oss-20B	-0.402954	0.000020	-0.312641	0.000016	-0.327418	0.000022	-0.269428	0.000022
gpt-oss-120B	-0.333963	0.000027	-0.199902	0.000024	-0.273421	0.000028	-0.161542	0.000031
gpt-5	0.156599	0.000025	0.299576	0.000026	0.218496	0.000026	0.318572	0.000029
Claude Sonnet 4	0.104098	0.000046	0.040367	0.000010	0.063612	0.000032	0.105911	0.000026
Claude Opus 4	0.347516	0.000065	0.171656	0.000061	0.221169	0.000048	0.332988	0.000067
Claude Opus 4.1	**0.452939**	**0.000146**	**0.470892**	**0.000141**	**0.445388**	**0.000141**	**0.493213**	**0.000203**
Qwen-3-0.6B	0.293593	0.000002	0.176909	0.000007	0.211447	0.000005	0.271553	0.000003
Qwen-3-1.7B	0.365808	0.000027	0.274632	0.000025	0.323745	0.000032	0.391681	0.000034
Qwen-3-4B	0.216582	0.000049	0.221653	0.000037	0.239971	0.000032	0.256920	0.000061
Qwen-3-8B	0.308547	0.000034	0.301039	0.000026	0.302231	0.000035	0.327433	0.000035
Qwen-3-14B	0.324731	0.000068	0.221253	0.000047	0.259408	0.000044	0.312000	0.000080
Qwen-3-32B	0.388342	0.000076	0.203790	0.000058	0.264353	0.000057	0.399234	0.000065
Qwen3-30B-A3B	0.329273	0.000050	0.385471	0.000074	0.372684	0.000043	0.416213	0.000069
Qwen3-235B-A22B	0.391492	0.000081	0.430555	0.000041	0.406943	0.000076	0.453300	0.000079
Deepseek-R1	-0.031810	0.000007	-0.045531	0.000004	-0.049256	0.000003	-0.010844	0.000002
V3.1	0.396647	0.000035	0.204811	0.000036	0.271420	0.000042	0.355928	0.000036
V3.1-Terminus	0.324070	0.000023	0.283478	0.000035	0.309142	0.000031	0.322799	0.000031
V3.2-Exp	0.302922	0.000027	0.265136	0.000033	0.272557	0.000020	0.321917	0.000029

Table 4: Performance of mainstream models in mathematical and scientific reasoning. Each benchmark shows ARISE scores and corresponding Scaling Metrics (SM). V3.1, V3.1-Terminus, and V3.2-Exp are all DeepSeek series models. The Scaling Metric shows relatively small values with the first three digits typically being zeros.

Model	Art & Design	Business	Science	Health & Med.	Human. & Soc.	Tech & Eng.	Avg.
GPT-5	**0.272**	**0.298**	**0.312**	**0.289**	**0.265**	0.274	**0.285**
o3	0.238	0.252	0.261	0.245	0.232	0.242	0.245
o4-mini	0.192	0.205	0.208	0.198	0.189	0.196	0.198
Sonnet 4	0.158	0.172	0.175	0.165	0.156	0.164	0.165
Opus 4	0.225	0.239	0.245	0.232	0.218	0.233	0.232
Opus 4.1	0.268	0.285	0.295	0.278	0.262	**0.280**	0.278

Table 5: Discipline-level ARISE analysis across MMMU domains. We evaluate six state-of-the-art reasoning models across six core disciplines. **Bold** values indicate the best performance in each column. The results demonstrate consistent cross-discipline scaling patterns, with Science-related domains generally achieving higher ARISE scores.

Model	SWE-bench Verified		LiveCodeBench		τ2-Bench		BFCL-v3	
	ARISE	SM	ARISE	SM	ARISE	SM	ARISE	SM
o1	0.120531	0.000048	0.104464	0.000025	0.097844	0.000016	0.090596	0.000019
o3	0.242285	0.000041	0.227926	0.000033	0.188289	0.000025	0.163869	0.000019
o3-mini	0.169071	0.000030	0.147321	0.000027	0.106837	0.000025	0.116984	0.000025
o4-mini	0.196871	0.000037	0.191874	0.000027	0.169686	0.000035	0.170538	0.000033
gpt-oss-20B	-0.326036	0.000023	-0.336262	0.000016	-0.384107	0.000007	-0.374236	0.000009
gpt-oss-120B	-0.209097	0.000025	-0.234670	0.000030	-0.240233	0.000017	-0.238353	0.000029
gpt-5	0.239273	0.000022	0.261642	0.000026	0.174972	0.000012	0.233048	0.000021
Claude Sonnet 4	0.092875	0.000033	0.067902	0.000030	0.067379	0.000028	0.064168	0.000021
Claude Opus 4	0.305950	0.000053	0.289900	0.000054	0.232681	0.000024	0.229159	0.000035
Claude Opus 4.1	0.480285	0.000133	0.483796	0.000129	0.439709	0.000078	0.431055	0.000079
Qwen-3-0.6B	0.270375	0.000001	0.213543	0.000004	0.170712	0.000001	0.208552	0.000005
Qwen-3-1.7B	0.334988	0.000033	0.305905	0.000028	0.291700	0.000018	0.286917	0.000022
Qwen-3-4B	0.230608	0.000039	0.247055	0.000031	0.189426	0.000022	0.182345	0.000025
Qwen-3-8B	0.306297	0.000040	0.301868	0.000026	0.271688	0.000023	0.283367	0.000020
Qwen-3-14B	0.289547	0.000064	0.269849	0.000048	0.236769	0.000039	0.277388	0.000033
Qwen-3-32B	0.350025	0.000056	0.339766	0.000055	0.235912	0.000077	0.292756	0.000049
Qwen3-30B-A3B	0.385370	0.000051	0.387557	0.000067	0.304281	0.000056	0.344193	0.000041
Qwen3-235B-A22B	0.446157	0.000049	0.430845	0.000062	0.389649	0.000068	0.370387	0.000062
Deepseek-R1	-0.028327	0.000003	-0.012076	0.000010	-0.035678	0.000007	-0.051757	0.000002
DeepSeek-V3.1	0.358224	0.000040	0.331861	0.000045	0.278489	0.000018	0.292173	0.000029
DeepSeek-V3.1-Terminus	0.329678	0.000021	0.310457	0.000029	0.260189	0.000025	0.297023	0.000020
DeepSeek-V3.2	0.316112	0.000029	0.282013	0.000040	0.246096	0.000030	0.256818	0.000029

Table 6: Performance of mainstream models on code generation and agentic benchmarks. Each benchmark shows ARISE scores and corresponding Scaling Metrics (SM). DeepSeek-V3.1, V3.1-Terminus, and V3.2 refer to the DeepSeek series models.