2025-03-10-文献汇报

> ⓘ Info
>
> **论文标题**：《s1: Simple test-time scaling》
>
> **作者**：Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, Tatsunori Hashimoto
>
> **arXiv地址**：https://arxiv.org/abs/2501.19393
>
> **作者团队**：斯坦福、华盛顿大学（西雅图）、艾伦人工智能研究所、Contextual AI
>
> GitHub：https://github.com/simplescaling/s1

v1(https://arxiv.org/abs/2501.19393v1) Fri, 31 Jan 2025 18:48:08 UTC (812 KB)
v2(https://arxiv.org/abs/2501.19393v2) Mon, 3 Feb 2025 16:31:30 UTC (804 KB)
[v3] Sat, 1 Mar 2025 06:07:39 UTC (810 KB)

研究背景

传统范式：

- 训练时间扩展 (Train Time Scaling Law)：通过扩展参数规模、数据大小和计算算力，大模型的能力会出现显著提升（关注数据/参数规模与模型性能）
- 存在问题：
  - 计算成本指数增长（GPT-4训练成本预估达6300万美元）
  - 边际效益递减（MATH数据集上GPT-4比GPT-3.5仅提升11.2%）

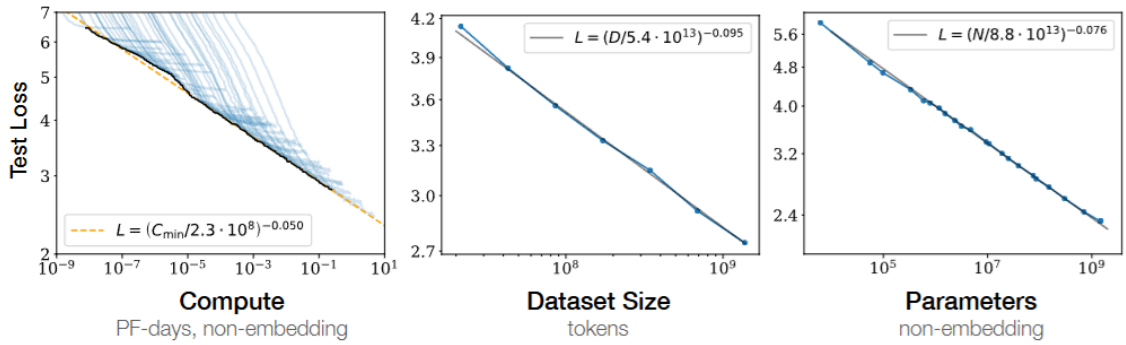


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

新范式探索：

- 测试时间扩展（Test Time Scaling Law）：增加推理时间能够显著提升 LLM 的推理质量（关注 token 与模型性能）
- **早期尝试**：
  - OpenAI o1（2024）：首提"推理时计算扩展"概念，但未公开技术细节
  - DeepSeek-R1（2025）：用800K数据+RLHF复现o1性能，失去简易性优势

> **Info**
>
> 在大语言模型（LLMs）的发展历程中， Scaling Laws 一直是推动性能提升的核心策略。研究表明，随着模型规模和训练数据的增长，LLMs 的表现会不断优化。然而，随着训练阶段规模的进一步扩大，性能提升的边际收益逐渐减小，训练更强大的模型需要巨额投入。因此，**研究重点逐渐从训练阶段的扩展转向推理阶段的扩展**，探索在不增加模型参数量的情况下，如何提升推理质量。
>
> 「慢思考」（Slow-Thinking），也被称为测试时扩展（Test-Time Scaling），成为提升 LLM 推理能力的新方向。近年来，OpenAI 的 o1 、DeepSeek 的 R1 以及 Qwen 的 QwQ 等顶尖推理大模型的发布，进一步印证了推理过程的扩展是优化 LLM 逻辑能力的有效路径。研究发现，增加推理时间能够显著提升 LLM 的推理质量，这一发现推动了对 「慢思考」方法的深入研究。

1. **研究问题**：这篇文章研究了如何在测试时通过增加计算资源来提高语言模型的推理性能。具体来说，作者探讨了最简单的实现测试时扩展和强大推理性能的方法。
2. **研究难点**：如何在有限的训练数据下实现高效的推理模型，以及如何有效地控制测试时的计算资源以获得最佳性能。

贡献：

1. **简单的方法来创建高效的推理数据集**：作者提出了一种简单的方法来创建一个高效的小型推理数据集（s1K），该数据集包含1,000个经过精心挑选的问题及其推理过程和答案。通过使用难度、多样性和质量三个标准进行筛选，确保了数据集的有效性。
2. **测试时扩展方法**：作者开发了一种简单的测试时扩展方法，称为预算强制（budget forcing），通过在模型生成过程中强制终止或延长思考时间来控制测试时的计算量。这种方法能够显著提高模型的推理性能。
3. **模型构建与性能**：基于上述数据集和方法，作者构建了一个名为s1-32B的模型，该模型在数学推理任务上表现出色，特别是在竞赛数学问题上超过了OpenAI的o1-preview模型，显示出强大的推理能力。

研究方法

构建 s1K 推理数据集

- **数据集选择原则：**
  - **质量 (Quality)：** 保证问题格式良好、推理链条清晰。
  - **难度 (Difficulty)：** 选取具有一定挑战性的问题，要求较多推理步骤。
  - **多样性 (Diversity)：** 涉及数学、物理、化学、生物等多个领域，覆盖不同推理任务。
- **数据收集与筛选步骤：**

1. **数据收集**：初始数据集包括从16个来源收集的59,029个问题，涵盖数学、物理、化学等多个领域。通过Google Gemini Flash Thinking API生成推理轨迹和解决方案。涵盖多个现有数据集（例如 NuminaMATH、OlympicArena、AGIEval 等）以及自创数据集（如 s1-prob 和 s1-teasers）

```
- s1-prob由斯坦福大学统计系博士资格考试（https：//statistics.stanford.edu）概率部分的182个问题组成，并附有涵盖困难证明的手写解决方案。概率资格考试每年举行一次，要求解决专业水平的数学问题。
- s1-teasers包括23个具有挑战性的脑筋急转弯，通常用于量化交易职位的面试问题。每个样本都由取自PuzzledQuant(https://www.puzzledquant.com/)的一个问题和解决方案组成。我们只取难度最高的例子（“难”）。
```

2. **样本选择**：最终选择了1000个问题作为训练数据，这些样本结合了难度、多样性和质量三个标准。具体步骤包括：

```
- 去除API错误和格式问题的样本。(质量)
- 使用两个模型评估每个问题的难度，去除过于简单的样本。（难度）
- 按领域分类问题，并从每个领域中随机选择一个样本，直到达到1000个样本。（多样性）
```

Table 6. **Summary of our dataset s1K**. Token count measured by the Qwen-2.5 tokenizer. We prompt Claude to produce keywords given several questions from the domain.

Domain	#questions	Total token count	Keywords
Geometry	109	560.2K	Area, Triangle, Distance
Number theory	98	522.5K	Sequences, Divisibility
Combinatorics	75	384.7K	Permutations, Counting
Real functions	43	234.8K	Trigonometry, Calculus
Biology	41	120.9K	Organic reactions
Complex functions	32	170.2K	Complex roots
Quantum theory	32	127.9K	Particles, Wave functions
Field theory	28	150.1K	Polynomials, Roots
Calculus of variations	28	155.5K	Optimization, Control
Difference equations	24	132.5K	Recurrence, Recursion
Electromagnetic theory	23	95.8K	Optics, Waves, Diffraction
Group theory	22	100.0K	Groups, Automorphisms
Linear algebra	22	128.3K	Matrices, Determinants
Probability theory	20	114.6K	Random walk, Expectation
Algebraic systems	19	109.9K	Functional equations
Mechanics	19	103.6K	Forces, Motion, Energy
Thermodynamics	19	74.2K	Heat engines, Entropy
Differential equations	18	89.6K	Substitution, Existence
Computer science	18	34.2K	Complexity theory, Algorithms
Numerical analysis	18	76.5K	Error analysis, Stability
Calculus	17	96.3K	Convergence, Summation
Algebraic structures	17	90.4K	Inequalities, Sets
Astronomy	16	37.7K	Stellar populations, Orbits
Remaining 27 domains	242	982.2K	Domains with ≤ 16 questions
All domains (51)	1000	4.7M	**s1K**

预算强制

通过简单地附加思考结束令牌分隔符和可选的 “Final Answer:” 来实施最大 token 计数，以提前退出思考阶段，并使模型提供其当前的最佳答案。

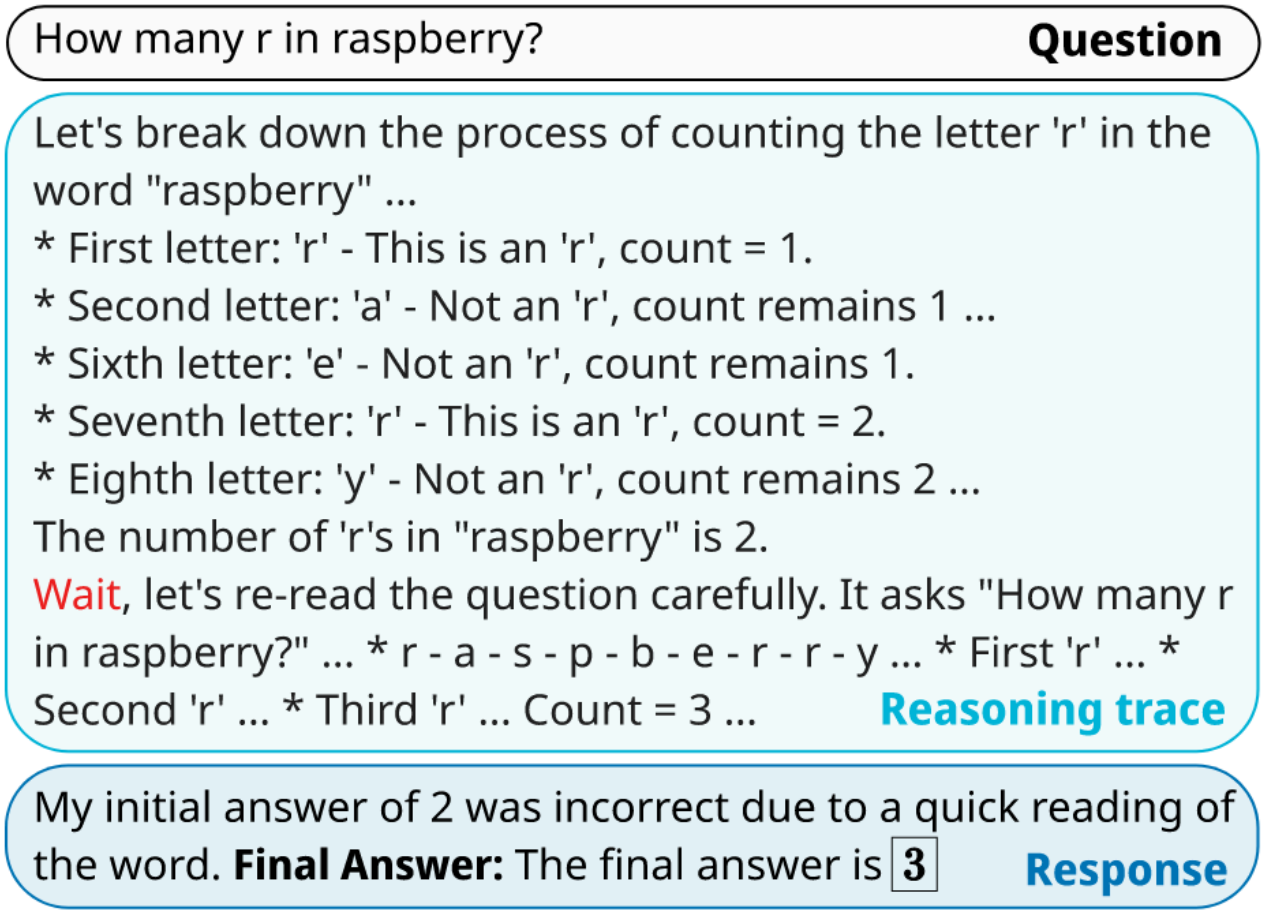


Figure 3. **Budget forcing with s1-32B.** The model tries to stop after "...is 2.", but we suppress the end-of-thinking token delimiter instead appending "Wait" leading **s1-32B** to self-correct its answer.

Baselines: 我们用以下方法对预算强制进行基准测试：

- (I) 条件长度控制方法，该方法依赖于在提示中告诉模型它应该生成多长时间。我们按粒度将它们分组为
  - (a) Token 条件控制：我们在提示中指定思维Token的上限；
  - (b) Step 条件控制：我们指定思维步骤的上限，其中每个步骤在100个 token 左右；
  - (c) Class 条件控制：我们编写两个通用提示，告诉模型思考短时间或长时间
- (II) 拒绝采样，其采样直到一代符合预定计算预算。这个预言捕获了以其长度为条件的后验over响应。

Results

Setup

训练: 在 Qwen2.5-32B 上执行监督微调——使用s1K数据集训练获得模型s1-32B。使用PyTorch FSDP 在16个 Nvidia H100 GPU 上进行微调花了26分钟。

评估任务:

- MATH500：数学竞赛问题。
- AIME24：美国数学邀请赛问题。
- GPQA Diamond：博士级科学问题。

基准测试:
OpenAI o1系列（OpenAI，2024），popularized test-time scaling 的闭源模型；

- DeepSeek r1系列（DeepSeekAI等人，2025），性能高达o1级的开放权重推理模型；
- Qwen的QwQ-32B-preview（团队，2024），一个没有公开方法的32B开放权重推理模型；
- Sky-T1-32B-Preview（团队，2025）和 Bespoke-32B（实验室，2025），开放模型，具有从QwQ-32B-preview 和 r1 中蒸馏的开放推理数据；
- Google Gemini 2.0 Flash Thinking Experimental（Google，2024），我们从中提取的API。由于它没有官方的评估分数，我们使用Gemini API自己对其进行基准测试。然而，Gemini API的"背诵错误"（recitation error）使评估变得具有挑战性。2我们通过在错误不会出现的web界面中手动插入所有30个AIME24问题来规避这一点。然而，我们省略了MATH500（500个问题）和GPQA Diamond（198个问题），因此它们在表1中不适用。
- 我们的模型s1-32B是完全开放的，包括权重、推理数据和代码。

Performance

测试时扩展效果:

结果表明，s1-32B模型在使用预算强制方法后，性能随着测试时计算资源的增加而提升

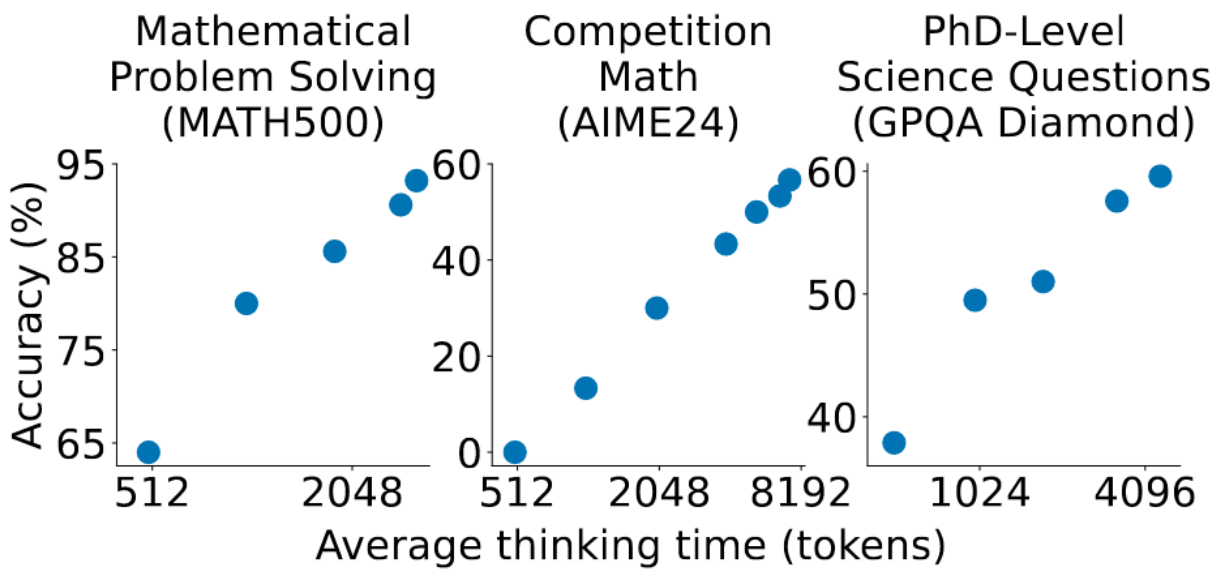
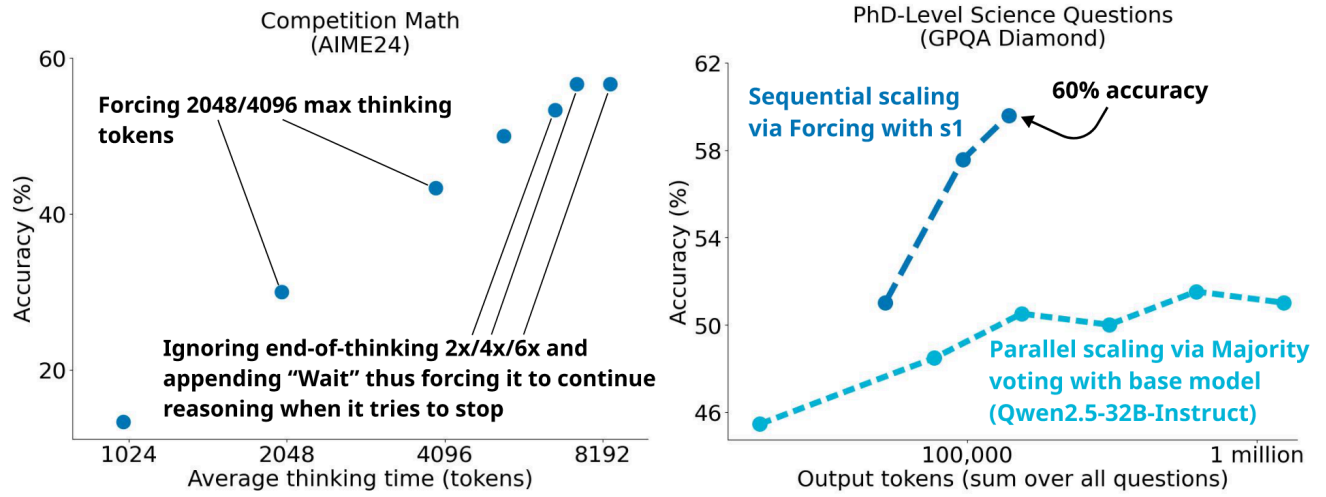


Figure 1. **Test-time scaling with s1-32B.** We benchmark **s1-32B** on reasoning-intensive tasks and vary test-time compute.

Budget Forcing

在图4（左）中，我们扩展了图1（中）的图，显示虽然我们可以使用我们的预算强制技术（§3）和更多的测试时间计算来提高AIME24性能，但它最终会在六倍时变平。过于频繁地抑制思维结束标记分隔符会导致模型进入重复的循环，而不是持续的推理。

在图4（右）中，我们展示了在我们的1,000个样本上训练Qwen2.5-32B-Instruct 以产生 s1-32B 并为其配备简单的预算强制技术后，与并行扩展（如多数投票）相比，顺序生成的扩展方法（即 Budget Forcing）表现更佳，因为后续生成能依赖于前面的中间结果进行迭代修正



(a) Sequential scaling via budget forcing

(b) Parallel scaling via majority voting

Figure 4. **Sequential and parallel test-time scaling.** *(a):* Budget forcing shows clear scaling trends and extrapolates to some extent. For the three rightmost dots, we prevent the model from stopping its thinking 2/4/6 times, each time appending "Wait" to its current reasoning trace. *(b):* For Qwen2.5-32B-Instruct we perform 64 evaluations for each sample with a temperature of 1 and visualize the performance when majority voting across 2, 4, 8, 16, 32, and 64 of these.

- 微调后的 s1-32B 模型（只使用 1000 个样本进行训练）在样本效率上表现突出，与闭源的 o1-preview 模型基本持平甚至部分指标高出近 27%。
- 对比采用全 59K 样本训练的情况，虽然全数据微调也能取得较好效果，但资源成本高很多，展现了 s1K 数据集筛选的重要性。

结果对比

在图2（右）和表1中，我们将s1-32B与其他型号进行了比较。我们发现s1-32B是样本效率最高的开放数据推理模型。尽管只在额外的1,000个样本上训练它，但它的性能明显好于我们的基本模型（Qwen2.5-32B-Instruct）。同时发布的r1-32B显示出比s1-32B更强的性能，同时也仅使用SFT（DeepSeek-AI等人，2025）。然而，它是在800倍以上的推理样本上训练的。仅仅用1000个样本是否能达到他们的性能是一个悬而未决的问题。最后，我们的模型几乎与AIME24上的 Gemini 2.0 Thinking 相匹配。

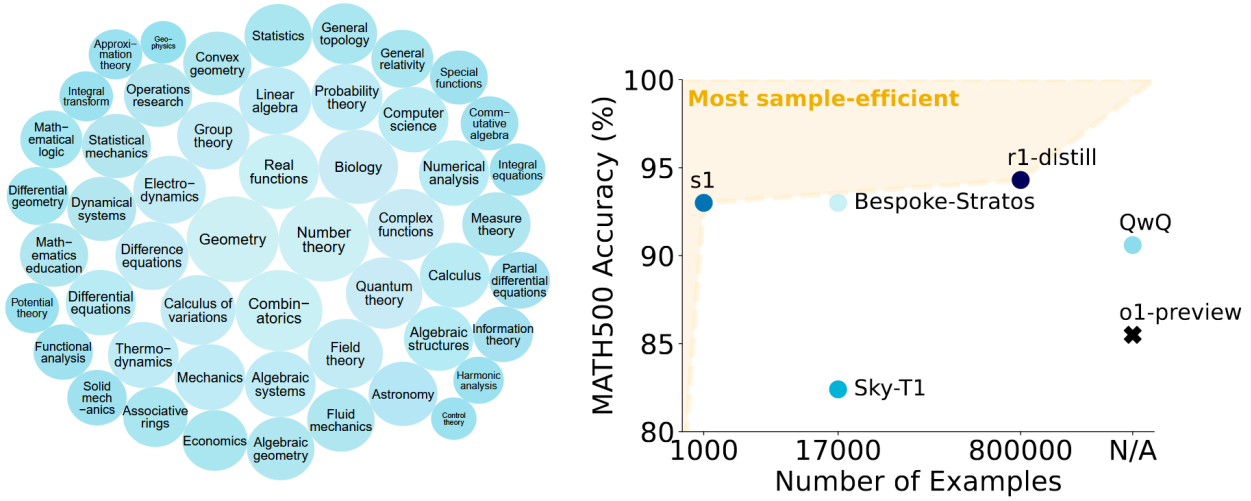


Figure 2. **s1K and s1-32B.** *(left)* **s1K** is a dataset of 1,000 high-quality, diverse, and difficult questions with reasoning traces. *(right)* **s1-32B**, a 32B parameter model finetuned on **s1K** is on the sample-efficiency frontier. See Table 1 for details on other models.

Table 1. **s1-32B is a strong open reasoning model.** We evaluate **s1-32B**, Qwen, and Gemini (some entries are unknown (N.A.), see §4). Other results are from the respective reports (Qwen et al., 2024; Team, 2024; OpenAI, 2024; DeepSeek-AI et al., 2025; Labs, 2025; Team, 2025). # ex. = number examples used for reasoning finetuning; BF = budget forcing. See §A for our better **s1.1** model.

Model	# ex.	AIME 2024	MATH 500	GPQA Diamond
API only				
o1-preview	N.A.	44.6	85.5	73.3
o1-mini	N.A.	70.0	90.0	60.0
o1	N.A.	**74.4**	**94.8**	**77.3**
Gemini 2.0 Flash Think.	N.A.	60.0	N.A.	N.A.
Open Weights				
Qwen2.5-32B-Instruct	N.A.	26.7	84.0	49.0
QwQ-32B	N.A.	50.0	90.6	54.5
r1	≫800K	**79.8**	**97.3**	**71.5**
r1-distill	800K	72.6	94.3	62.1
Open Weights and Open Data				
Sky-T1	17K	43.3	82.4	56.8
Bespoke-32B	17K	**63.3**	93.0	58.1
s1 w/o BF	**1K**	50.0	92.6	56.6
s1-32B	**1K**	56.7	**93.0**	**59.6**

Ablations

数据筛选重要性：

- 仅随机选取 1000 个样本或仅保证多样性/难度单一指标均会显著降低模型性能。
- 三者联合（质量、难度、多样性）筛选得到数据集 s1K，能在较小样本下实现推理性能的大幅提升。

Table 2. **s1K data ablations.** We budget force (BF) a maximum of around 30,000 thinking tokens for all scores in this table. This performs slightly better than the scores without BF (Table 1) as it allows the model to finish with a best guess when stuck in an infinite loop. We report 95% paired bootstrap confidence intervals for differences relative to the **s1K** model using 10,000 bootstrap samples. E.g., the interval [-13%, 20%] means that, with 95% confidence, the true difference between 59K-full and **s1K** is between -13% and +20%. If the entire interval is negative, e.g. [-27%, -3%], we can confidently say that the performance is worse than **s1K**.

Model	AIME 2024	MATH 500	GPQA Diamond
1K-random	36.7 [-26.7%, -3.3%]	90.6 [-4.8%, 0.0%]	52.0 [-12.6%, 2.5%]
1K-diverse	26.7 [-40.0%, -10.0%]	91.2 [-4.0%, 0.2%]	54.6 [-10.1%, 5.1%]
1K-longest	33.3 [-36.7%, 0.0%]	90.4 [-5.0%, -0.2%]	59.6 [-5.1%, 10.1%]
59K-full	53.3 [-13.3%, 20.0%]	92.8 [-2.6%, 2.2%]	58.1 [-6.6%, 8.6%]
s1K	50.0	93.0	57.6

测试时扩展控制对比：

- 分别尝试了基于token 数（token-conditional）、基于步骤数（step-conditional）以及基于类别（class-conditional）的控制方法。
- 通过 Budget Forcing 干预，能够实现 100% 的控制精度，从而使得生成过程在限制内停止，同时展现明确的性能提升趋势。
- 拒绝采样方法在较短的生成长度下表现较好，但在较长的生成长度下出现反向扩展趋势

Table 3. **Ablations on methods to scale test-time compute on AIME24.** $|\mathcal{A}|$ refers to the number of evaluation runs used to estimate the properties; thus a higher value indicates more robustness. **Bold** indicates our chosen method and the best values. BF = budget forcing, TCC/SCC/CCC = token/step/class-conditional control, RS = rejection sampling.

Method	Control	Scaling	Performance	$\lvert\mathcal{A}\rvert$
BF	**100%**	15	**56.7**	5
TCC	40%	-24	40.0	5
TCC + BF	**100%**	13	40.0	5
SCC	60%	3	36.7	5
SCC + BF	**100%**	6	36.7	5
CCC	50%	**25**	36.7	2
RS	**100%**	-35	40.0	5

总结

Important

这篇论文提出了一种简单的方法来实现测试时扩展和强大的推理性能。通过创建一个小型数据集s1K并使用预算强制技术，作者成功地训练了一个高效的推理模型s1-32B。该模型在多个基准测试中表现出色，并且在测试时扩展方面具有显著优势。论文的贡献包括：

1. 开发了创建高效推理数据集和测试时扩展的简单方法。
2. 基于这些方法构建了s1-32B模型，并与o1-preview模型竞争。
3. 通过消融实验验证了数据集选择和测试时扩展方法的有效性