Emergent Abilities in Reasoning Models during Inference-Time Scaling

Anonymous Authors[1]

Abstract

The emergence of new abilities under train-time scaling (more data or parameters) is a well-known phenomenon, yet it remains unclear whether similar abrupt transitions occur under test-time scaling (longer reasoning). In the present work, we study whether gains in performance (and confidence) are smooth or exhibit phase-transition-like jumps as the available token budget during generation increases. To quantify emergence robustly, we introduce two metrics: Breakthroughness^{+} (the largest gain across token budgets, normalised by the average of all gains) and Burstiness (the degree to which total gains are concentrated in a small number of budget increases). We evaluate three reasoning-model families spanning 1.5B–32B parameters on mathematical (AIME 24 and 25) and scientific (GPQA) benchmarks. Across settings, we observe three consistent patterns: (1) emergence is markedly sharper on maths than on science, where scaling is smoother; (2) the same examples very consistently exhibit non-linear scaling across models; and (3) larger-size models exhibit emergence more frequently than smaller-size ones. Put together, our results indicate that emergent behaviours under test-time scaling are predictable from a complex combination of task, example, and model properties.

1. Introduction

Inference-time compute scaling improves the accuracy of large language models (LLMs) in reasoning-intensive tasks. This is facilitated by prompting and post-training strategies that increase accuracy at the expense of efficiency at inference time. By producing long reasoning traces before reporting a final answer (Wei et al., 2022b) or sampling multiple reasoning traces first before aggregating them (Wang et al., 2023), models such as OpenAI o1 (OpenAI et al., 2024) and DeepSeek-R1 (Guo et al., 2025) have shown dramatic improvements on a wide range of benchmarks, including mathematical, scientific, and coding problems.

Previously, model scaling was mostly focused on the training phase: Kaplan et al. (2020) and Hoffmann et al. (2022) (*inter alia*) established "scaling laws" that can reliably link model size, training dataset size, and training compute to pre-training loss. These provide engineering and safety benefits by allowing developers to anticipate computational requirements and model risks, without committing significant resources. However, studies have also identified "emergent abilities" in LLMs—specific tasks exhibiting sharp, abrupt improvements when scaling exceeds a certain threshold (Wei et al., 2022a; Srivastava et al., 2022; Ganguli et al., 2022; Berti et al., 2025). For instance, Chinchilla and Gopher model families may remain at random-chance level performance in some MMLU tasks below 10B parameters, but improve suddenly beyond them (Hendrycks et al., 2020; Wei et al., 2022a; Hoffmann et al., 2022; Rae et al., 2021).

While similar scaling laws have been verified for inference-time scaling, too—finding a power law relationship between generated token budget and model accuracy (Snell et al., 2024; Muennighoff et al., 2025; Brown et al., 2025)—to the best of our knowledge, possible emergent behaviours during inference-time scaling still remain to be studied systematically. In fact, the aforementioned laws may be an artefact of aggregating scaling behaviour across a wide number of tasks, which may smooth out non-linear trends that occur at a more granular level. Crucially, the possible existence of emergent abilities puts into question the extent to which the effects of additional inference-time compute (including potential model risks) are consistent and predictable, nullifying the benefits of scaling laws.

Hence, this study examines cases in which sharp, abrupt increases in model performance occur as inference-time compute budget is scaled—"Aha!" moments in reasoning traces (Gandhi et al., 2025) that resemble sudden moments of insight in human cognition (Kounios & Beeman, 2009)—and the key factors that influence this behaviour. The first challenge that this endeavour faces is properly quantifying emergence with appropriate metrics. While research on training-time emergence emphasised the need for more

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

expressive (Alabdulmohsin et al., 2022) or "broken" (Caballero et al., 2023) scaling laws to better fit scaling data, these do not directly measure emergence and do not apply to individual examples. Moreover, previously proposed emergence metrics (Srivastava et al., 2022) behave inconsistently across models (see Section 2). The second challenge is that recorded emergence may be a "mirage" of discrete evaluation metrics, which disappears once considering continuous ones (Schaeffer et al., 2023).

As a solution, our first contribution is to define new metrics that capture emergence at the level of either individual examples or tasks and behave robustly across models. These consist of Breakthroughness$^+$ and Burstiness, which represent normalised measures of the size of the largest gain across inference-time compute budgets and the dispersion of such gains. In addition, we apply these metrics to both discrete metrics (i.e., accuracy) and continuous ones (i.e., probability and negentropy). Our second main contribution is a comprehensive empirical analysis of emergent abilities during inference-time scaling. This includes multiple families and scales of reasoning models—DeepSeek-R1-Distill 1.5B / 7B / 14B / 32B (Guo et al., 2025), QwQ-32B (QwenTeam, 2025), and Phi-4-Reasoning-Plus (14B) (Abdin et al., 2025), multiple datasets—AIME (24 and 25) for maths and GPQA-Diamond for science—and computing budgets from 2^7 to 2^{13} generated tokens. The resulting metrics allow us to address a series of research questions:

RQ1 Does the degree of emergence during inference-time scaling vary across different tasks (e.g., maths vs. science)?
RQ2 Given our emergence metrics, do the same examples in each task consistently exhibit emergent scaling across multiple models?
RQ3 Are larger reasoning models more or less likely to display inference-time emergent behaviours compared to smaller reasoning models?

We find that not only are emergent behaviours pervasive in inference-time scaling, but also that their presence depends on a complex interaction of properties at the level of task, example features, and model size. We provide our code and metrics at `https://github.com/anonymised`.

2. Method

2.1. Scaling inference-time compute

This study focuses on sequential scaling of inference-time compute, as opposed to parallel scaling or hybrid approaches. This choice is made as it appears to be the most natural setting in which models can exhibit sudden breakthroughs in performance analogous to "Aha!" moments of insight displayed by humans (Kounios & Beeman, 2009).[1] To control the length of a reasoning trace, we use the budget forcing method of Muennighoff et al. (2025), appending special tokens to force a model to end or continue its generation. To make a model output a reasoning trace and answer of length T tokens (the token budget), we begin by sampling from the model in standard fashion; if the model tries to end its response before hitting the token budget by returning `<EOS>`, we remove `<EOS>` and append `Hmm, let me keep thinking` to force the model to reconsider and continue working. Once the model nears the token budget T, we append `The final answer is:` to force the model to output a solution, allowing 10 tokens to provide a solution, and then end generation.

2.2. Datasets and models

We study scaling behaviour on mathematical and science benchmarks, due to the proficiency of language reasoning models (LRMs) on these tasks. We use the American Invitational Mathematics Examination (AIME) 2024 and 2025 datasets, which contain 30 challenging mathematical problems each with integer solutions in the range 0-999, and GPQA-Diamond, a challenging subset of 448 multiple-choice questions testing domain expertise in biology, physics, and chemistry.

These datasets are chosen due to their closed solution sets—the AIME datasets have integer solutions in the range 0-999, and GPQA has multiple choice answers in the set $\{A, B, C, D\}$. This allows for renormalisation of the model output distribution over the solution sets, which yields more insightful metrics to track over inference compute budgets (discussed in the next section).

We study emergent inference-time scaling behaviour on three popular open-weight reasoning models:

- DeepSeek-R1-Distill-32B (Guo et al., 2025)
- QwQ-32B (QwenTeam, 2025)
- Phi-4-Reasoning-Plus (14B) (Abdin et al., 2025)

Additionally, we also investigate scaling behaviour with respect to model size by evaluating against the DeepSeek-R1-Distill family—specifically the 1.5B, 7B, 14B, and 32B models (Guo et al., 2025).

[1] More precisely, sampling multiple responses from a model appears unlikely to replicate the settings in which humans display sudden breakthrough moments in reasoning—instead, this comes from reconsideration and revising of previous attempts, which is closer to sequential scaling.

2.3. Metrics

The benchmarks have closed solution sets: AIME-24 and AIME-25 have ground truths $y \in \{0, 1, 2, \ldots, 999\}$ and GPQA has $y \in \{A, B, C, D\}$, allowing the model's output distribution to be renormalised over the solution set $\mathcal{S}$. Letting M denote a reasoning model that outputs reasoning trace r and answer y^* to input prompt x, the model output is evaluated over three metrics:

Accuracy: Binary correctness score

$$s = \mathbf{1}[y^* = y]$$

Probability of Ground Truth: Model's assigned probability to the correct answer, renormalised over the solution set

$$p_{\text{gt}} = \frac{p_M(y|x,r)}{\sum_{s\in\mathcal{S}} p_M(s|x,r)}$$

Negative Entropy: Negative entropy of the renormalised distribution over all candidate solutions

$$-H = -\sum_{s\in\mathcal{S}} \tilde{p}(s) \log \tilde{p}(s)$$

where $\tilde{p}(s) = \frac{p_M(s|x,r)}{\sum_{s'\in\mathcal{S}} p_M(s'|x,r)}$.

Accuracy is discrete, whereas probability of ground truth and negative entropy are continuous; all three metrics are tracked to determine if emergent scaling behaviour appears only for discrete metrics, as suggested in Schaeffer et al. (2023), or across a wider range of measures. Note that negative entropy measures a model's confidence over the solution set, not performance directly. A model can have high negative entropy but low accuracy/ground truth probability if confident in the incorrect answer—a trend that can be observed in some of the empirical results (Section 3).

2.4. Scoring emergent scaling

This study introduces two metrics for scoring the degree to which scaling behaviour across the three metrics above is sharp and abrupt. The first, Breakthroughness$^+$ improves upon a similar metric proposed in (Srivastava et al., 2022), which measures the ratio of the total change in response variable y to the average change between two consecutive points. The second, Burstiness, is the Gini coefficient of the absolute differences between consecutive points in the scaling curve.

In this section we briefly define Breakthroughness$^+$ and Burstiness, with further technical details in Section B.

2.4.1. Breakthroughness$^+$

Breakthroughness$^+$ is defined as:

$$\text{Breakthroughness}^+ = \frac{(\max_i y_i - \min_i y_i)^2}{\text{MeanSquareRoot}(\{\Delta y_i\}_i)} \quad (1)$$

The ratio between the total change in response variable y ($\max_i y_i - \min_i y_i$) to the average change between two consecutive points in the scaling curve (MeanSquareRoot($\{\Delta y_i\}_i$)) is the core component of the metric; this captures the extent to which a single jump in y dominates the overall scaling curve. We use the `MeanSquareRoot`[2] to capture the average change between two consecutive points in the scaling curve–this operator dampens the effect of outlier jumps (see Section B.1 for more details of this choice). An additional factor multiplicative factor, ($\max_i y_i - \min_i y_i$), is added to the numerator to make the metric responsive to large jumps in y; without this factor, the metric would not distinguish between response variable vectors that jump from 0 to 1, and from 0 to 100. Note that the Breakthroughness$^+$ score is only defined for positive values of Δy_i due to the use of the `MeanSquareRoot` operator. We use an additional emergence score, Burstiness, to compensate for this limitation.

2.4.2. Burstiness

Burstiness is defined as:

$$\text{Burstiness} = \underbrace{(\max_i y_i - \min_i y_i)}_{w} \times \underbrace{\frac{\sum_i \Delta_i}{\sum_i |\Delta_i|}}_{R} \times \underbrace{\text{Gini}(|\Delta y_i|)}_{G} \quad (2)$$

Here, G is the Gini coefficient of the absolute differences between consecutive points, which captures the abruptness of the response variable scaling. R captures the directionality of the change (abrupt increase or decreases in the response variable), and w makes the metric responsive to the magnitude of change. The Gini coefficient is used in economic studies to measure the inequality of wealth distributions; in our case, we adapt it to measure the concentration of the cumulative change in the response variable to a small number of jumps. Readers who seek more details on the Gini coefficient are referred to Section B.1.1, which also contains illustrative examples of the metric applied to a range of different scaling behaviours.

For now, we note that both metrics capture heuristic notions of emergent scaling, rather than being derived formally, hence using two metrics provides greater robustness to the presence of the scaling behaviour that we are interested in. If both metrics identify a task as exhibiting strong emergent scaling, this provides stronger evidence for the presence of the behaviour than reliance on a single metric.

[2] Defined as $\texttt{MeanSquareRoot}(\{\Delta y_i\}_i) = \left(\frac{1}{N-1}\sum_{i=1}^{N-1}\sqrt{\Delta y_i}\right)^2$.

3. Results

The following section presents the results of the study. We begin with an investigation of emergent inference-time scaling trends across different datasets—GPQA, AIME24, and AIME25. Following this, we study emergent inference-time scaling across individual problem instances within AIME25, to see if the same instances consistently display emergent scaling across models. Finally, we examine trends across four sizes of DeepSeek-R1-Distill (1.5B, 7B, 14B, and 32B), to see if larger reasoning models are more likely to exhibit sharp, abrupt increases in performance.

3.1. Trends across datasets

Do all tasks exhibit smooth scaling with respect to additional inference-time compute, or do some show a propensity for more erratic scaling behaviour? It is this question that we answer in this section. We investigate this question using DeepSeek-R1-Distill-Qwen-32B, with results for Phi-4-Reasoning-Plus and QwQ-32B shown in Section C.2.

Distributions of Breakthroughness$^+$ and Burstiness scores obtained by evaluating the scaling profiles of each instance across GPQA, AIME25, and AIME24 are shown in Figure 1, with 95% confidence intervals for population means (calculated via bootstrapping with $N = 1000$ samples) displayed in the legend. Distributions for the ground truth probability are shaded in blue, whereas distributions for negative entropy over the solution sets are shaded in orange to emphasise the conceptual difference between the two[3].

A visual inspection of the emergence score distributions suggests that AIME24 and AIME25 exhibit significantly sharper inference-time scaling than GPQA; GPQA distributions across all combinations of metric-emergence score pairings are significantly more concentrated at low values relative to the AIME datasets. Mann-Whitney U tests for statistical significance are presented in Section C.1, and find a significant difference between GPQA and the AIME datasets for all combinations of metric-emergence score pairings, but no significant difference between AIME24 and AIME25. This is further supported by observing that the 95% population mean confidence intervals between GPQA and the AIME datasets show minimal overlap.

The aggregate scaling curves on the right of Figure 3 display the implications of these distributional differences on scaling behaviour; GPQA exhibits smooth returns to additional inference-time compute for all metrics (accuracy, ground truth probability, and negative entropy), whereas the AIME tasks see sharp increases in all three metrics at roughly 1024 thinking tokens. It is worth noting that AIME24 exhibits sharper scaling than AIME25, rising from 0% to $\sim$ 60% accuracy after the critical 1024 budget mark, whilst AIME25 increases from 0% to $\sim$ 40%. However, we do not attribute this to distributional differences between AIME24 and AIME25 tasks; given that all models in this study were released in 2025, we expect this to be the result of training dataset contamination with AIME24.

What might be the reason for the difference in emergent inference-time scaling between the AIME datasets and GPQA? One explanatory hypothesis is related to the compositional complexity of tasks—there are theoretical reasons to believe that tasks that are composed of many cleanly separable subproblems are more likely to exhibit abrupt scaling (Arora & Goyal, 2023a; Barak, 2023). We explore this hypothesis and related ones later in Section 4.

3.2. Instance analysis

Having shown that some datasets exhibit sharp aggregate improvements, we now investigate whether emergent scaling depends on individual problem features rather than dataset- or model-level properties. Specifically, we test whether the same problems consistently exhibit emergent inference-time scaling across different models. Our analysis focuses on AIME25 as it exhibited strong emergent scaling at the aggregate level (see Section 3.1), with corresponding results for AIME24 and GPQA shown in Section D.

Figure 4 shows Spearman correlations for the emergence scores of AIME25 instances between different models. Each problem instance is assigned a ranking based on its Breakthroughness$^+$ and Burstiness score[4], and correlation coefficients calculated between these rankings. Between all model pairings, a strong correlation coefficient is observed (between $0.88 - 0.89$), providing strong evidence for the hypothesis that *across all reasoning models, the same problem instances consistently exhibit emergent inference-time scaling.*. That is, features of individual problem instances have a strong influence on the sharpness of inference-time scaling behaviour.

Figure 5 shows scaling curves that support this conclusion whilst also adding important nuance. This figure shows scaling behavior of accuracy, ground truth probability, and negative entropy for the top-3 instances exhibiting emergent inference-time scaling when evaluated **with DeepSeek-R1-Distill**, across all models (i.e, DeepSeek-R1-Distill, Phi-4-Reasoning-Plus, and QwQ-32B). The corresponding input prompts are shown in Table 1. It can be observed that scaling

[3]namely, that emergent inference-time scaling of ground truth probability represents a sudden increase in the model's confidence in the correct answer, whereas for negative entropy this reflects a sudden increase in the model's confidence of a solution, which may or may not be the correct one

[4]Specifically, each instance i receives a rank for its Breakthroughness$^+$ score and a rank for its Burstiness score, with the final rank being the average of the two.

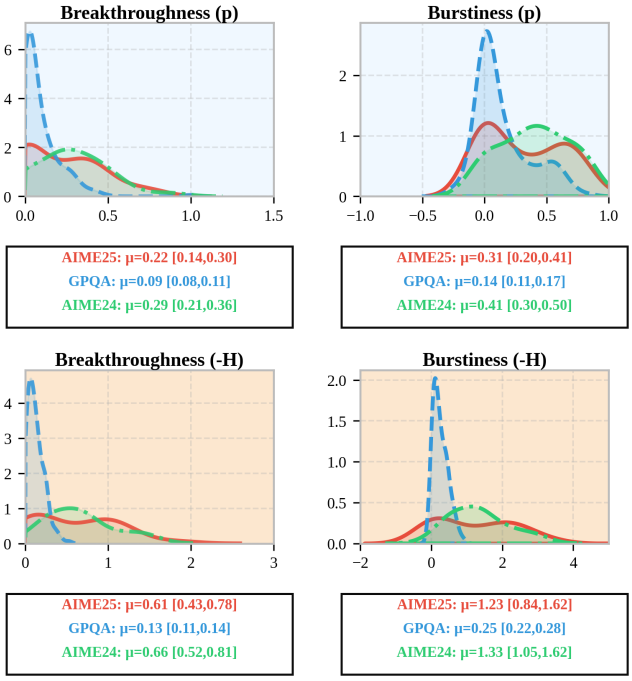


Figure 1

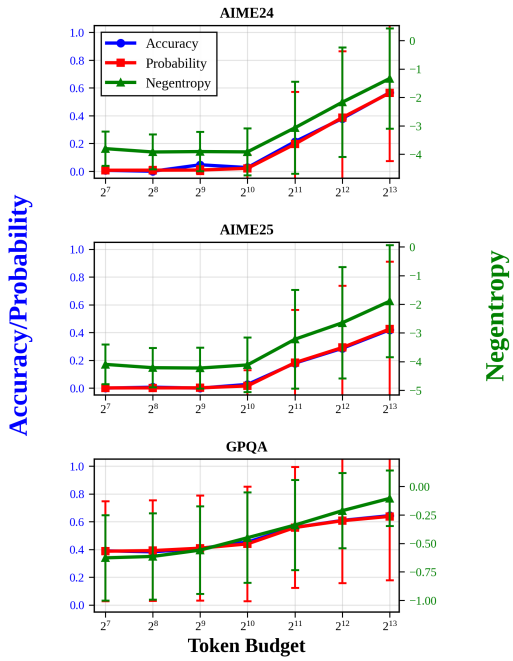


Figure 2

Figure 3. **Left**: Distributions of Breakthroughness$^+$ and Burstiness scores for ground truth probability and negative entropy across AIME24, AIME25, and GPQA. GPQA distributions are concentrated at low values indicating smooth inference-time scaling relative to AIMEs. Mann-Whitney U tests in Section C.1 confirm this difference is significant. **Right**: Aggregate scaling curves for AIME24, AIME25, and GPQA.AIME datasets shown an abrupt increase in all metrics at roughly 1024 thinking tokens, whereas GPQA exhibit smooth returns to inference-time compute across all toekn budgets. Results for Deepseek-R1-Distill-Qwen-32B, with results for Phi-4-Reasoning-Plus and QwQ-32B shown in Section C.2.

behaviour is similar across all reasoning models; to illustrate, within a single token budget doubling, instance 2 shows ¿75% increases in accuracy and ground truth probability across all models. *Importantly, however, the token budget at which this sharp transition occurs varies across models*; the abrupt jump happens between 1024 and 2048 thinking tokens for DeepSeek-R1-Distill and Phi-4-Reasoning-Plus, whereas for QwQ-32B, the jump happens between 2048 and 4096 thinking tokens. A similar trend holds for other instances, with Phi-4-Reasoning-Plus showing the earliest transitions in general, followed by DeepSeek-R1-Distill, and then QwQ-32B. We delay further discussion of this phenomenon to Section 4.

3.3. Trends across model size

Having studied models with similar parameter counts, we now investigate the relationship between model size and emergent inference-time scaling. Specifically, we evaluate four sizes of DeepSeek-R1-Distill (1.5B, 7B, 14B, and 32B) on the AIME25 dataset, with corresponding results for AIME24 shown in Section E.2. An analysis for GPQA is omitted as it does not exhibit strong emergent scaling on DeepSeek-R1-Distill 32B.

Instance Index	Problem
5	An isosceles trapezoid has an inscribed circle tangent to each of its four sides. The radius of the circle is 3, and the area of the trapezoid is 72. Let the parallel sides of the trapezoid have lengths r and s, with $r \neq s$. Find r^2+s^2.
2	The 9 members of a baseball team each had a single scoop cone of chocolate, vanilla, or strawberry ice cream. At least one player chose each flavor, and the number of players who chose chocolate > vanilla > strawberry. Let N be the number of assignments; find the remainder when N is divided by 1000.
15	Points A, B, C, D, E, F lie in order on a line. G is off the line. $AC = 26$, $BD = 22$, $CE = 31$, $DF = 33$, $AF = 73$, $CG = 40$, $DG = 30$. Find the area of $\triangle BGE$.

Table 1. AIME25 top-3 instances displaying emergent inference-time scaling

Figure 7 shows summary statistics for the emergence score distributions plotted against model size, with uncertainty intervals representing the standard error of the mean. From

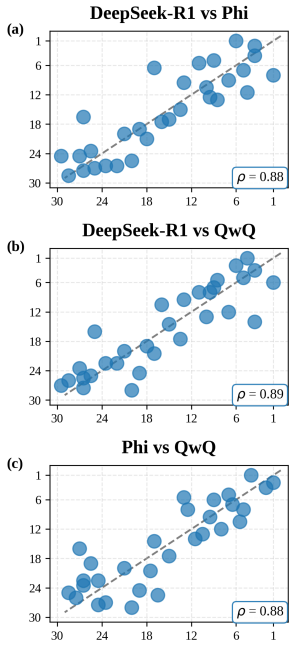


Figure 4

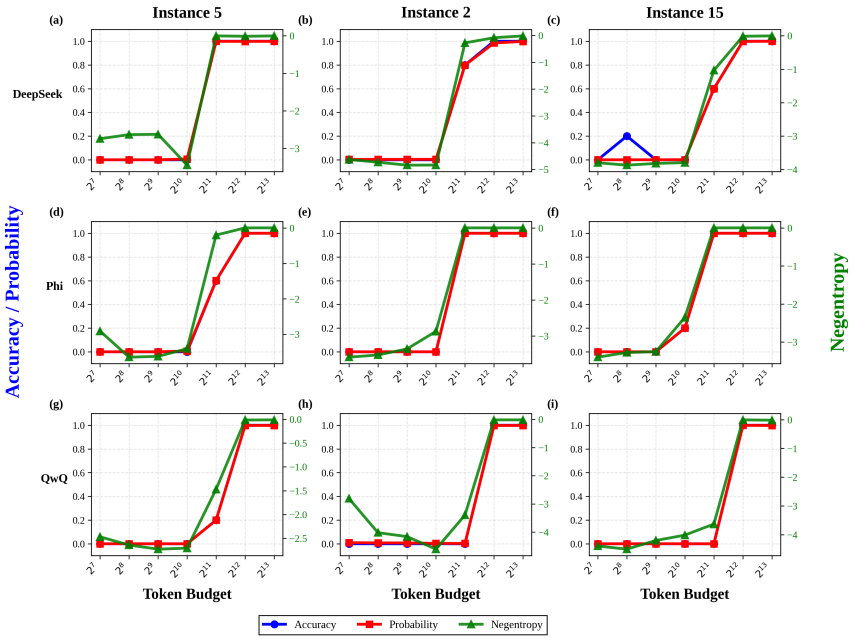


Figure 5

Figure 6. **Left**: Spearman correlations for the emergence scores of AIME25 instances between different models. Between all model pairings, a strong correlation coefficient is observed (between 0.88 – 0.89), indicating that problem features are strong predictors of emergent scaling, rather than model-level properties. **Right**: Scaling curves for top-3 AIME25 instance showing emergent scaling validates dependence on problem features, but highlight an important caveat: the onset of emergent scaling can be influenced by the reasoning model.

this, we observe an increasing trend in emergent inference-time scaling across model size when tracking ground truth probability (blue background), with a weaker trend appearing for the negative entropy metric (orange background). Further statistical analysis with Mann-Whitney U tests, and Cohen's d finds a medium-to-large difference between the emergence scores of the 1.5B model and the 7B/14B/32B models ($d > 0.5$), and a small difference between the 7B and 14B/32B models ($0.2 < d < 0.5$). Negligible differences in emergent inference-time scaling are observed between the 14B and 32B models ($d \approx 0.1$).

Aggregate scaling curves are plotted in Figure 8 that corroborate these conclusions. The 1.5B model exhibits a small increase in performance around the 2048 thinking token threshold, rising to 20% accuracy at 8096 tokens; on the other hand, the 14B and 32B models exhibit a sharp increase in performance at 1024 tokens, attaining final accuracies of 35% and 40% respectively. The 7B model sits in between, seeing a smoother increase in performance beginning at 1024 tokens and rising to a final accuracy of 25% at 8096 tokens.

Interestingly, negative entropy shows weaker trends across model sizes, with the 14B model exhibiting the sharpest increases, suggesting a weak relationship between model size and solution confidence.

Taken together, this evidence suggests that sharp, abrupt increases in model performance are more likely in larger reasoning models than smaller ones. An important implication of this is that methods for predicting abilities of smaller models when using additional inference-time compute may not scale to larger counterparts, which is discussed further in Section 4.

4. Discussion

In this section, we discuss the implications of the three main findings of Section 3 in further detail.

There is significant variation in the degree of emergent inference-time scaling across different tasks, with mathematical reasoning datasets (AIME24, AIME25) displaying strong emergent scaling behaviour whilst scientific reasoning and recall tasks (GPQA) exhibit much smoother returns to inference-time compute. AIME problems may require algorithms needing a certain number of serial steps to complete, mapping to the 1024 thinking tokens, after which model performance increases sharply. Insufficient thinking budgets would result in failed algorithm execution and lower performance.

Differences in compositional complexity between AIME and GPQA problems could also explain emergent inference-time

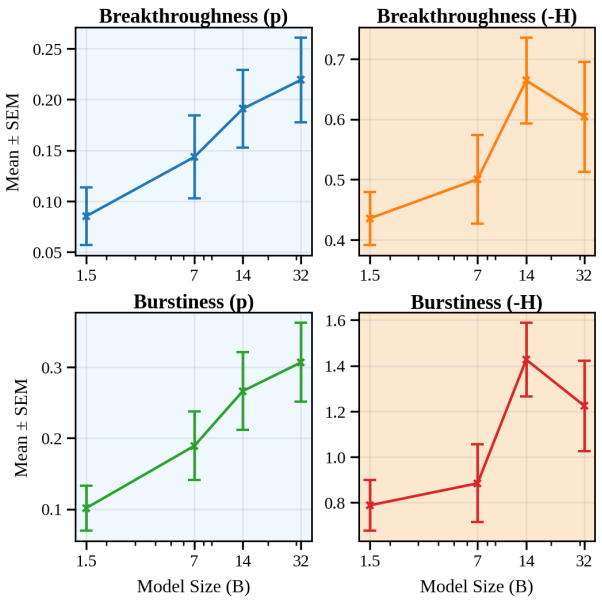


Figure 7. Summary statistics (Mean ± Standard Error) for the emergence score distributions (Breakthroughness$^+$ and Burstiness evaluated on ground truth probability and negative entropy scaling profiles) against model size. Blue background indicates ground truth probability, orange background indicates negative entropy. For ground truth probability, central statistics are generally increasing across model size, suggesting that larger models are more likely to show sharp improvements in performance as inference compute is scaled. Trends are less clear for negative entropy, with the 14B model showing the sharpest negative entropy scaling; although 14B and 32B model show significant differences from the 1.5B and 7B models. Results for AIME24 are given in Section E.2.

scaling. The compositional complexity hypothesis, a popular explanation for emergent abilities during training (Arora & Goyal, 2023a; Barak, 2023; Okawa et al., 2024), could manifest at inference-time as follows: up until the 1024 token budget, the model solves subproblems (proving intermediate lemmas, computing partial results), then composes these into a final answer. GPQA problems may lack this compositional structure, exhibiting smoother scaling. Further exploration of these hypotheses provides a clear direction for future research.

Additionally, this result shows the linear relationship between the logarithm of inference-time compute and model performance is not always valid, with two-piece piecewise linear curves being more appropriate for the AIME datasets in this study. In the training domain, similar fits have already been proposed to extend simple linear fits (Alabdulmohsin et al., 2022; Caballero et al., 2023) ("broken neural scaling laws"); extending this to more datasets in the inference-time domain seems like a natural direction for future work.

Emergent inference-time scaling is strongly influenced by features of individual examples, with the same instances across AIME25 (as well as AIME24 and GPQA), consistently exhibiting emergent scaling across multiple models. However, an important caveat applies: the *onset* of emergent inference-time scaling (the token budget at which the sharp increase in performance occurs) can vary across models, with Phi-4-Reasoning-Plus exhibiting the earliest transitions in general, followed by DeepSeek-R1-Distill, and then QwQ-32B. Therefore, analyses of emergent scaling that focus on the magnitude and abruptness of the change in performance should focus on instance-level properties, such as compositional complexity, solution space complexity, and number of serial steps required, whilst the token budget at which emergent scaling occurs should account for model-level properties, such as the structure of reasoning traces (Marjanović et al., 2025). On the latter point, we can speculate that Phi-4-Reasoning-Plus is more concise and direct in its reasoning process, whilst QwQ-32B engages in broader exploration over the solution space before committing to an answer, which could explain its delayed onset of emergent scaling. Therefore, we are eager to see future studies that investigate systematic differences in reasoning methods across different models.

Larger models are more likely to exhibit emergent inference-time scaling than smaller models across performance metrics (accuracy, ground truth probability), but not confidence metrics (negative entropy). The key implication of this finding is that extrapolating inference-time scaling laws found for smaller models to larger models may not be valid, as the scaling behaviour of larger models may follow a different functional form altogether. This also suggests that the (inference-time) scaling behaviour of larger models is naturally more unpredictable than smaller models; this motivates additional care when deploying larger models in production environments relative to smaller ones. Finally, it is important to note that the trend of increased emergent scaling for larger models was only weakly observed for negative entropy. Arguably, this is an undesirable property; DeepSeek-R1-Distill 1.5B showed similar negative entropy scaling profiles to the 32B model, despite achieving a final accuracy of 20% compared to 40% for the 32B model. Ideally, less competent models would reflect this by exhibiting lower confidence in their final solutions, but our finding implies that incorrect models are as confident in incorrect solutions as correct models are in the right ones. Resolving this issue is another avenue that future work could take.

5. Conclusion

This study investigated the presence of sharp, abrupt increases in model performance as inference-time compute is scaled, previously referred to as "emergent abilities" when studied in the training domain. This was motivated by evidence in the training domain that log-linear scaling re-

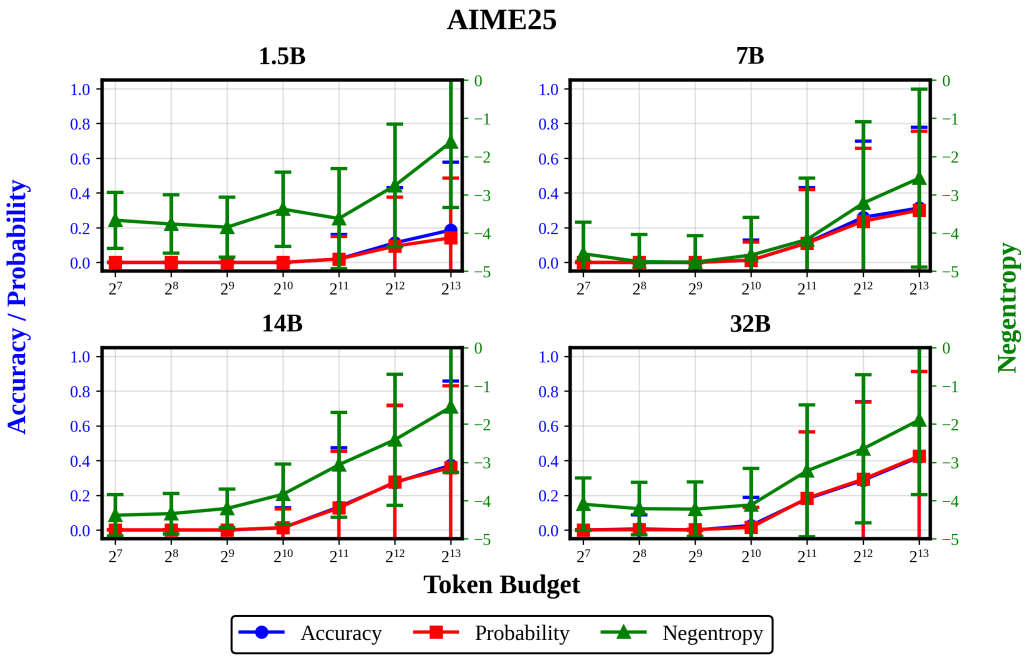


Figure 8. Aggregate scaling behaviour for AIME25 across four model sizes (1.5B, 7B, 14B, and 32B). The 1.5B model shows smoother scaling than then larger models (14B, 32B), and only achieves non-zero accuracy at 4096 tokens, compared to 2048 tokens for the other model sizes. The 7B model sees relatively gradual returns to additional inference-time compute past 1024 tokens, whilst the 14B and 32B models display a noticeably steeper gradient. Note that the negative entropy scaling profiles across models shows less variation than the accuracy and ground truth probability profiles, with th 1.5B model reaching similar levels of negative entropy as the 32B model at the final budget of 8192 tokens. The 14B model achieves the lowest negative entropy of the four sizes, reaching ~ -1 bits at the final budget.

lationships are sometimes inadequate, and the notion that models may exhibit sudden breakthrough moments during reasoning, akin to "Aha!" moments of insight displayed by humans (Kounios & Beeman, 2009). Model performance and confidence were tracked using discrete and continuous metrics over reasoning budgets of up to 8192 tokens, and identified with two new metrics for emergent scaling – Breakthroughness^{+} and Burstiness.

The analysis revealed three main findings—the first being that mathematical reasoning tasks exhibit substantially stronger emergent scaling than scientific reasoning and recall tasks, with a two-piece piecewise linear curve being a better model for this scaling behaviour than a linear fit. Additionally, instance-level features were found to be a key determinant of emergent scaling, with the same instances across all datasets consistently exhibiting sharp scaling behaviour, independent of the model used for evaluation. However, the *onset* of emergence—the token budget at which transitions occur—varies across models, suggesting that model-level properties such as reasoning trace structure also play a role. Finally, larger models were found to exhibit more abrupt scaling behaviour than smaller models when tracking performance metrics, with a weaker trend occurring when tracking confidence metrics (negative entropy).

These results motivate an investigation into the mechanisms that lead to emergent inference-time scaling. Emergent inference-time scaling could be explained by numerous hypotheses, such as a compositional complexity hypothesis, which already has reasonable theoretical backing (Arora & Goyal, 2023a; Barak, 2023). More speculatively, emergent inference-time scaling may occur when models make sudden breakthroughs in reasoning traces, cases of which have previously been identified in the DeepSeek-R1 model (Guo et al., 2025) and human cognition (Kounios & Beeman, 2009). Future work should explore whether abrupt improvements in performance as inference compute increases are localised to a small number of influential tokens. Should emergent inference-time scaling present itself across an even wider range of tasks and models, we should expect that new, undiscovered abilities of language models can be unlocked by scaling inference compute even further.

References

Abdin, M., Agarwal, S., Awadallah, A., Balachandran, V., Behl, H., Chen, L., de Rosa, G., Gunasekar, S., Javaheripi, M., Joshi, N., Kauffmann, P., Lara, Y., Mendes, C. C. T., Mitra, A., Nushi, B., Papailiopoulos, D., Saarikivi, O., Shah, S., Shrivastava, V., Vineet, V., Wu, Y., Yousefi, S., and Zheng, G. Phi-4-reasoning technical report, 2025. URL https://arxiv.org/abs/2504.21318.

Alabdulmohsin, I., Neyshabur, B., and Zhai, X. Revisiting neural scaling laws in language and vision, 2022. URL https://arxiv.org/abs/2209.06640.

Arora, S. and Goyal, A. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023a.

Arora, S. and Goyal, A. A theory for emergence of complex skills in language models, 2023b. URL https://arxiv.org/abs/2307.15936.

Barak, B. Emergent abilities and grokking: Fundamental, mirage, or both? https://windowsontheory.org/2023/12/22/emergent-abilities-and-grokking-fundamental-mirage-or-both/, 2023. Blog post, Windows on Theory (accessed: 2024-06-01).

Berti, L., Giorgi, F., and Kasneci, G. Emergent abilities in large language models: A survey, 2025. URL https://arxiv.org/abs/2503.05788.

Brown, B., Juravsky, J., Ehrlich, R. S., Clark, R., Le, Q. V., Re, C., and Mirhoseini, A. Large language monkeys: Scaling inference compute with repeated sampling, 2025. URL https://openreview.net/forum?id=0xUEBQV54B.

Caballero, E., Gupta, K., Rish, I., and Krueger, D. Broken neural scaling laws, 2023. URL https://arxiv.org/abs/2210.14891.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Costarelli, A., Allen, M., Hauksson, R., Sodunke, G., Hariharan, S., Cheng, C., Li, W., Clymer, J., and Yadav, A. Gamebench: Evaluating strategic reasoning abilities of llm agents. *arXiv preprint arXiv:2406.06613*, 2024.

Du, Z., Zeng, A., Dong, Y., and Tang, J. Understanding emergent abilities of language models from the loss perspective, 2025. URL https://arxiv.org/abs/2403.15796.

Feng, K., Zhao, Y., Liu, Y., Yang, T., Zhao, C., Sous, J., and Cohan, A. Physics: Benchmarking foundation models on university-level physics problem solving. *arXiv preprint arXiv:2503.21821*, 2025.

Feng, X., Wan, Z., Wen, M., McAleer, S. M., Wen, Y., Zhang, W., and Wang, J. Alphazero-like tree-search can guide large language model decoding and training, 2024. URL https://arxiv.org/abs/2309.17179.

Gandhi, K., Chakravarthy, A. K., Singh, A., Lile, N., and Goodman, N. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective STars. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=QGJ9ttXLTy.

Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D., Elhage, N., et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1747–1764, 2022.

Gema, A. P., Hägele, A., Chen, R., Arditi, A., Goldman-Wetzler, J., Fraser-Taliente, K., Sleight, H., Petrini, L., Michael, J., Alex, B., et al. Inverse scaling in test-time compute. *arXiv preprint arXiv:2507.14417*, 2025.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Huang, Y., Hu, S., Han, X., Liu, Z., and Sun, M. Unified view of grokking, double descent and emergent abilities: A perspective from circuits competition. *arXiv preprint arXiv:2402.15175*, 2024.

Jain, N., Han, K., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., and Stoica, I. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kounios, J. and Beeman, M. The aha! moment: The cognitive neuroscience of insight. *Current directions in psychological science*, 18(4):210–216, 2009.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

Lubana, E. S., Kawaguchi, K., Dick, R. P., and Tanaka, H. A percolation model of emergence: Analyzing transformers trained on a formal language, 2024. URL https://arxiv.org/abs/2408.12578.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback, 2023. *URL https://arxiv. org/abs/2303.17651*, 2023.

Marjanović, S. V., Patel, A., Adlakha, V., Aghajohari, M., BehnamGhader, P., Bhatia, M., Khandelwal, A., Kraft, A., Krojer, B., Lù, X. H., et al. Deepseek-r1 thoughtology: Let's think about llm reasoning. *arXiv preprint arXiv:2504.07128*, 2025.

Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., and Hashimoto, T. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Okawa, M., Lubana, E. S., Dick, R. P., and Tanaka, H. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task, 2024. URL https://arxiv.org/abs/2310.09336.

OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C. M., de Bourcy, C., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., Tsipras, D., Li, D., Oprica, D., Freeman, E., Zhang, E., Wong, E., Proehl, E., Cheung, E., Mitchell, E., Wallace, E., Ritter, E., Mays, E., Wang, F., Such, F. P., Raso, F., Leoni, F., Tsimpourlas, F., Song, F., von Lohmann, F., Sulit, F., Salmon, G., Parascandolo, G., Chabot, G., Zhao, G., Brockman, G., Leclerc, G., Salman, H., Bao, H., Sheng, H., Andrin, H., Bagherinezhad, H., Ren, H., Lightman, H., Chung, H. W., Kivlichan, I., O'Connell, I., Osband, I., Gilaberte, I. C., Akkaya, I., Kostrikov, I., Sutskever, I., Kofman, I., Pachocki, J., Lennon, J., Wei, J., Harb, J., Twore, J., Feng, J., Yu, J., Weng, J., Tang, J., Yu, J., Candela, J. Q., Palermo, J., Parish, J., Heidecke, J., Hallman, J., Rizzo, J., Gordon, J., Uesato, J., Ward, J., Huizinga, J., Wang, J., Chen, K., Xiao, K., Singhal, K., Nguyen, K., Cobbe, K., Shi, K., Wood, K., Rimbach, K., Gu-Lemberg, K., Liu, K., Lu, K., Stone, K., Yu, K., Ahmad, L., Yang, L., Liu, L., Maksin, L., Ho, L., Fedus, L., Weng, L., Li, L., McCallum, L., Held, L., Kuhn, L., Kondraciuk, L., Kaiser, L., Metz, L., Boyd, M., Trebacz, M., Joglekar, M., Chen, M., Tintor, M., Meyer, M., Jones, M., Kaufer, M., Schwarzer, M., Shah, M., Yatbaz, M., Guan, M. Y., Xu, M., Yan, M., Glaese, M., Chen, M., Lampe, M., Malek, M., Wang, M., Fradin, M., McClay, M., Pavlov, M., Wang, M., Wang, M., Murati, M., Bavarian, M., Rohaninejad, M., McAleese, N., Chowdhury, N., Chowdhury, N., Ryder, N., Tezak, N., Brown, N., Nachum, O., Boiko, O., Murk, O., Watkins, O., Chao, P., Ashbourne, P., Izmailov, P., Zhokhov, P., Dias, R., Arora, R., Lin, R., Lopes, R. G., Gaon, R., Miyara, R., Leike, R., Hwang, R., Garg, R., Brown, R., James, R., Shu, R., Cheu, R., Greene, R., Jain, S., Altman, S., Toizer, S., Toyer, S., Miserendino, S., Agarwal, S., Hernandez, S., Baker, S., McKinney, S., Yan, S., Zhao, S., Hu, S., Santurkar, S., Chaudhuri, S. R., Zhang, S., Fu, S., Papay, S., Lin, S., Balaji, S., Sanjeev, S., Sidor, S., Broda, T., Clark, A., Wang, T., Gordon, T., Sanders, T., Patwardhan, T., Sottiaux, T., Degry, T., Dimson, T., Zheng, T., Garipov, T., Stasi, T., Bansal, T., Creech, T., Peterson, T., Eloundou, T., Qi, V., Kosaraju, V., Monaco, V., Pong, V., Fomenko, V., Zheng, W., Zhou, W., McCabe, W., Zaremba, W., Dubois, Y., Lu,

Y., Chen, Y., Cha, Y., Bai, Y., He, Y., Zhang, Y., Wang, Y., Shao, Z., and Li, Z. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

QwenTeam. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36:55565–55581, 2023.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Steinhardt, J. Future ml systems will be qualitatively different. https://bounded-regret.ghost.io/future-ml-systems-will-be-qualitatively-different/, 2022.

Sutton, R. S., Barto, A. G., et al. Reinforcement learning. *Journal of Cognitive Neuroscience*, 11(1):126–134, 1999.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Wei, J. Emergent abilities of large language models. https://www.jasonwei.net/blog/emergence, 2022.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Wikimedia Commons. Economics gini coefficient2.svg, 2009. URL https://en.wikipedia.org/wiki/Gini_coefficient#/media/File:Economics_Gini_coefficient2.svg. [Online; accessed 21-December-2024].

Wu, T.-Y. and Lo, P.-Y. U-shaped and inverted-u scaling behind emergent abilities of large language models. *arXiv preprint arXiv:2410.01692*, 2024.

Wu, Y., Sun, Z., Li, S., Welleck, S., and Yang, Y. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models, 2025. URL https://arxiv.org/abs/2408.00724.

Zhang, D., Zhoubian, S., Hu, Z., Yue, Y., Dong, Y., and Tang, J. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772, 2024.

A. Related Work

A.1. Scaling inference-time compute in language reasoning models

There are a wide range of methods for leveraging compute at inference-time to enhance model outputs, which can be broadly clustered into sequential and parallel methods. Sequential methods increase the length of the reasoning trace that a model produces, such as by fine-tuning it to critique and revise its own answer (Muennighoff et al., 2025; Madaan et al., 2023), or producing a chain-of-thought when reasoning (Wei et al., 2022b). This leads to a refined output distribution relative to the base model output (Snell et al., 2024). In contrast, parallel methods sample multiple responses from a model, then aggregate these responses using a method such as majority voting or best-of-N sampling (Brown et al., 2025; Wang et al., 2022; Cobbe et al., 2021; Lightman et al., 2023). To select amongst multiple responses, an outcome-based reward model may be applied to the solutions to determine the best response (Cobbe et al., 2021), but a better, though more expensive, approach is to train a process-based reward model (Lightman et al., 2023) to score intermediate steps and direct methods such as beam search (Feng et al., 2024), lookahead search, or Monte Carlo Tree Search (MCTS) (Sutton et al., 1999). Using additional compute at inference-time can enhance a model's performance across numerous domains, but well-structured tasks that require multi-step reasoning and verification particularly benefit (Zhang et al., 2024; He et al., 2024; Jimenez et al., 2023; Jain et al., 2024; Costarelli et al., 2024; Rein et al., 2024; Feng et al., 2025).

Multiple studies posit an approximately linear relationship between the logarithm of inference-time compute and model performance (Muennighoff et al., 2025; Brown et al., 2025). However, performance saturates after a certain amount of budget scaling, and in some cases begins to deteriorate—this phenomenon is known as inverse scaling (Gema et al., 2025). Relatedly, Marjanović et al. (2025) finds that U-shaped scaling behaviour can be observed as reasoning trace length is scaled. Given the wide range of methods to scale inference-time compute, studies such as Snell et al. (2024) and Wu et al. (2025) have studied how to optimally allocate compute at inference-time, finding that question difficulty is a useful statistic for determining optimal scaling strategy, and that compute-optimal scaling strategies can outperform naive best-of-N methods by up to 4x.

A.2. Emergent abilities of large language models

The seminal work on emergent abilities in large language models (Wei et al., 2022a) characterises emergent abilities as follows: "An ability is emergent if it is not present in smaller models but is present in larger models. *Emergent abilities would not have been directly predicted by extrapolating a scaling law from small-scale models.*" Wei et al. (2022a); Ganguli et al. (2022) observe that a large number of tasks within the BIG-Bench (Srivastava et al., 2022) and MMLU (Hendrycks et al., 2020) benchmarks exhibit emergent scaling behaviour with respect to model parameters and/or training compute. For a comprehensive list of 137 abilities identified as emergent with respect to training inputs, see Wei (2022).

Schaeffer et al. (2023) makes a key contribution to the emergent scaling literature, claiming that emergent scaling is not a fundamental property of a model or task, but can instead be attributed to the metric used for evaluation. They show that by replacing discrete metrics with continuous metrics, which recognise and award partial progress towards a solution, emergent scaling largely disappears. However, follow-up work (Berti et al., 2025) has questioned this claim, identifying methodological issues in the definition of emergent scaling used and arguing that it was defined too strictly. Additionally, other works have identified tasks that exhibit sharp, non-linear scaling behaviour with respect to both discrete and continuous metrics (Wei et al., 2022a; Steinhardt, 2022; Du et al., 2025).

Motivated by these phenomena, a growing body of literature seeks to identify proximate causes and the underlying mechanisms of emergent scaling. One popular hypothesis is that emergent scaling results from a high degree of compositional complexity in a task (Arora & Goyal, 2023b), meaning that a model must learn several underlying skills before it can complete the task; when the final skill is learnt, model performance increases rapidly. This has been formalised mathematically (Arora & Goyal, 2023b) and experimentally studied in vision (Okawa et al., 2024) and natural language (Lubana et al., 2024) domains. Alternative explanations link emergent scaling to critical thresholds in pretraining loss (Du et al., 2025), grokking and deep double descent (Power et al., 2022; Nakkiran et al., 2021; Huang et al., 2024), and aggregation effects across task distributions of varying complexity (Wu & Lo, 2024). None of the proposed mechanisms are mutually exclusive or broadly accepted as the primary explanation; thus, this topic remains an active area of research.

B. Emergence Scores

B.1. Breakthroughness$^+$

Srivastava et al. (2022) propose the Breakthroughness metric to quantify the degree to which a model is able to learn a task *"only once it grows beyond a critical scale."*. Denoting independent and response variable data as $\{(x_i, y_i)\}_{i=1}^N$ ordered by compute budget x_i, and differences between consecutive points as $\Delta y_i = y_{i+1} - y_i$, the Breakthroughness score of Srivastava et al. (2022) is:

$$B = \frac{s \cdot (\max_i y_i - \min_i y_i)}{\text{RootMedianSquare}(\{\Delta y_i\}_i)} \tag{3}$$

where $s = \text{sign}(\arg\max_i y_i - \arg\min_i y_i)$ captures the directionality of the performance change.

Intuitively, this metric can be thought of as the ratio of the overall change in the response variable y to the average change between consecutive points. The average change between consecutive points is found with the root median square operator, which does not factor outliers into the average, unlike averages based on taking the mean. This choice makes the Breakthroughness score take large values when the total change in response variable y is accounted for by a small number of jumps—a key characteristic of emergent scaling.

However, the Breakthroughness score has *two key limitations*. First, we find the removal of outlier metric differences with the root median square operator to be too crude. This problem is apparent in cases where a metric remains at approximately 0 before jumping to a non-zero value (the scaling behaviour of ground truth probability often exhibits this behaviour)—in such cases, the root median square operator returns an average difference of 0, leading to extremely large and degenerate Breakthroughness scores. We address this problem by replacing the root median square averaging operator with the mean square root operator, defined for metric differences Δy_i in equation (4). This operator dampens rather than eliminates the influence of outlier differences, acting as a more practical averaging method than the root median square.

$$\texttt{MeanSquareRoot}(\{\Delta y_i\}_i) = \left(\frac{1}{N-1}\sum_{i=1}^{N-1}\sqrt{\Delta y_i}\right)^2 \tag{4}$$

An additional change to the Breakthroughness metric is required. To see why, first consider the three response variable vectors:

$$\begin{aligned} Y_1 &= [0, 0, 0, 0, 1], \\ Y_2 &= [0, 0, 0, 0.8], \\ Y_3 &= [0, 0, 0, 0, 0.6]. \end{aligned}$$

Clearly, the scaling behaviour of Y_1 is sharper than Y_2, which is sharper than Y_3. However, replacing the RootMedianSquare operator with the MeanSquareRoot **only** would lead to the Breakthroughness metric assigning equal values to these three datasets. In other words, the original Breakthroughness metric is responsive to the **abruptness** of the metric change, but not its **magnitude**, which we refer to as the problem of **scale invariance**.

To make the Breakthroughness metric responsive to the magnitude of the metric change, in addition to the abruptness of the change, we simply weigh the numerator by the total change in the response variable y—$y_{max} - y_{min}$.

Therefore, our updated breakthroughness metric, Breakthroughness$^+$, is defined as:

$$B^+ = \frac{(\max_i y_i - \min_i y_i)^2}{\text{MeanSquareRoot}(\{\Delta y_i\}_i)} \tag{5}$$

The direction of the metric change s has been removed as Breakthroughness$^+$ is only defined for positive values of Δy_i.

The Breakthroughness$^+$ score improves upon the Breakthroughness metric in Srivastava et al. (2022), but it is not perfect. The most significant limitation is that it is defined only when the response variable y is consistently increasing. Whilst this assumption holds for our study, it will not hold in the presence of inverse scaling effects. Additionally, the Breakthroughness$^+$

score conceptualises emergence as a large ratio between the overall change in response variable y and the average change between consecutive points. This is a reasonable proxy, but there are other, conceptually distinct, ways to capture this behaviour. Therefore, we also use the Burstiness score to identify emergent scaling behaviour.

B.1.1. Burstiness

The Burstiness metric has its basis in the Gini coefficient, which is often used to measure the inequality of (wealth) distributions. Mathematically, the Gini coefficient is defined as (half of) the *relative mean absolute difference*—that is, the expected absolute difference between all pairings in a dataset. Given a dataset with values $y_1, y_2, \ldots, y_n$ and mean μ, the Gini coefficient (G) is:

$$G = \frac{1}{2n^2\mu} \sum_{i=1}^{n} \sum_{j=1}^{n} |y_i - y_j|$$

The geometric definition of the Gini coefficient is a more intuitive definition of distribution inequality; given the Lorenz curve (cumulative distribution) and the line of perfect equality (the line $y = x$ representing a case where the distribution mass is allocated equally across the domain) the Gini coefficient is the ratio of the area between the line of perfect equality and CDF, and the area below the line of perfect equality. This definition is illustrated below:

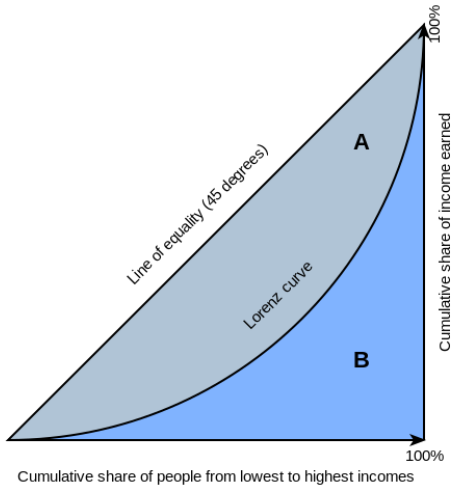


Figure 9. Geometric definition of the Gini (G) coefficient, where $G = \frac{A}{A+B}$ (when x is the cumulative fraction of population, and y is the cumulative share of wealth.). Figure from Wikimedia Commons Wikimedia Commons (2009).

However, the original Gini metric is defined for non-negative distributions, requiring some modification for our use case. Our proposed Burstiness metric first computes the Gini coefficient of the distribution of absolute differences $|\Delta y_i|$, which represents the degree to which the change in response variable y is concentrated in a small number of jumps. Figure 10 presents the absolute difference distribution and Gini coefficients for curves of interest.

For linear curves (top row), the absolute difference distribution (third column) is uniform, leading to a linear cumulative density function of the differences (fourth column). Applying the Gini coefficient formula, this leads to a Gini coefficient of approximately 0. On the other hand, the Heaviside function (second row)—a characteristic case of emergent scaling—has a delta function derivative; when these differences are sorted, the area between the cumulative distribution (purple line, fourth column) and line of perfect equality (dotted line) reaches it's maximum value of 0.5, leading to a Gini coefficient of ~ 1.0. Intermediate cases between these two extreme examples are given in the rest of the figure, where it is clear that the smoother exponential function has a more evenly distributed absolute difference distribution than the step functions, resulting in a smaller Gini coefficient. Note also how the metric adapts for curves with negative differences, such as the Multi-step function (row 4).

The Burstiness metric has two additional factors. A multiplicative term, $R = \frac{\sum_i \Delta_i}{\sum_i |\Delta_i|}$ (with $R \in [-1, 1]$) is added to capture the direction of the variable change. Additionally, the Gini coefficient is scale-invariant, meaning that it accounts

only for the abruptness of the variable change, and not the magnitude. To account for this, we add the weighting factor $w = \max_i y_i - \min_i y_i$ that was previously used to make the Breakthroughness+ metric responsive to scale.

Therefore, the Burstiness metric (B) is the product:

$$\text{Burstiness} = \underbrace{(\max_i y_i - \min_i y_i)}_{w} \times \underbrace{\frac{\sum_i \Delta_i}{\sum_i |\Delta_i|}}_{R} \times \underbrace{\text{Gini}(|\Delta y_i|)}_{G} \tag{6}$$

where G is the Gini coefficient of $|\Delta y_i|$ capturing the abruptness of a variable change, R captures the net direction of change, and w makes the metric responsive to the magnitude of the change.

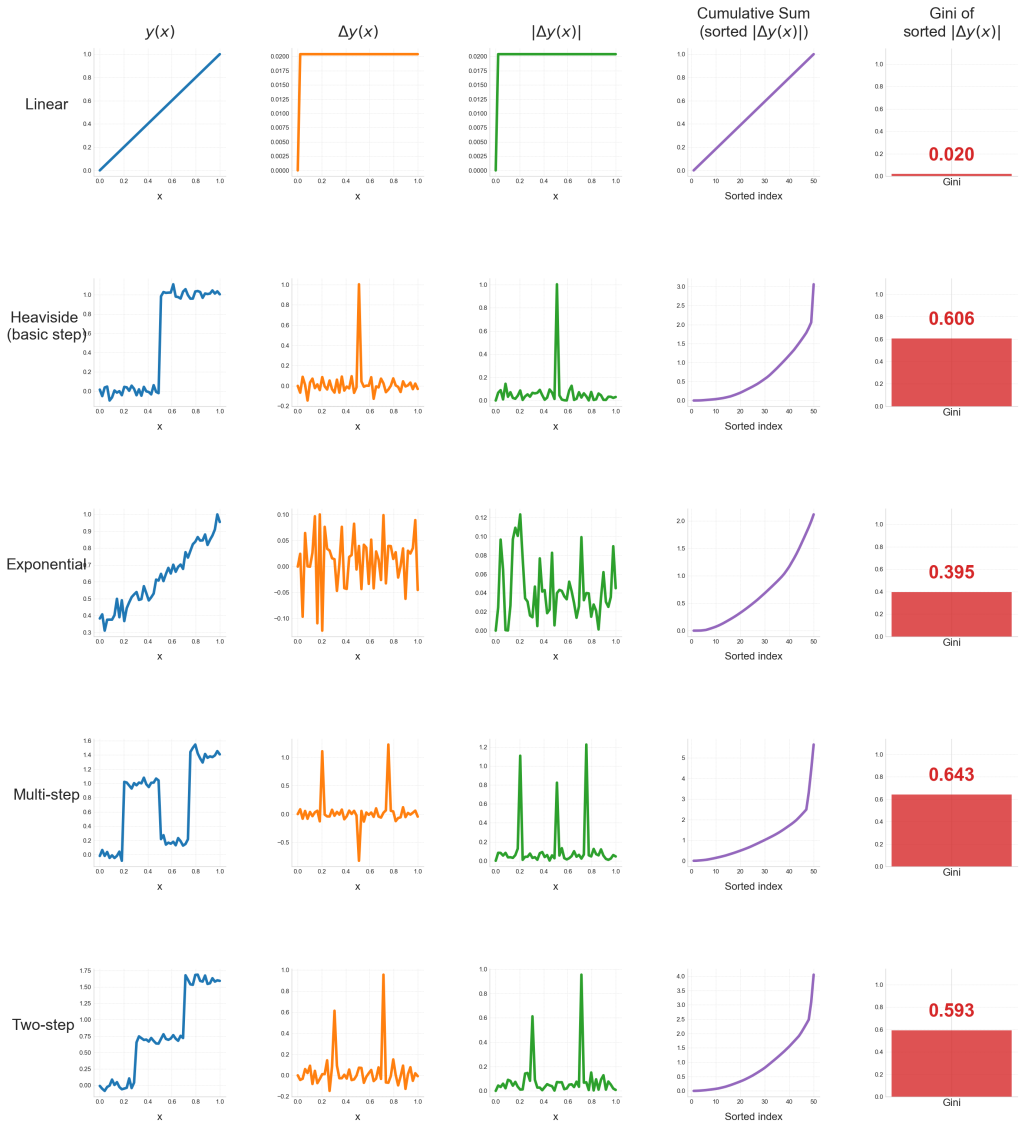


Figure 10. Gini coefficient of absolute differences for scaling profiles of interest. Note that this is **not** the Burstiness metric, which multiplies $G(|\Delta y_i|)$ by w and R.

C. Dataset Trends - Additional Analysis

C.1. Additional analysis for Deepseek-R1 (32B)

Mann-Whitney U tests for statistical significance are presented in the table below.

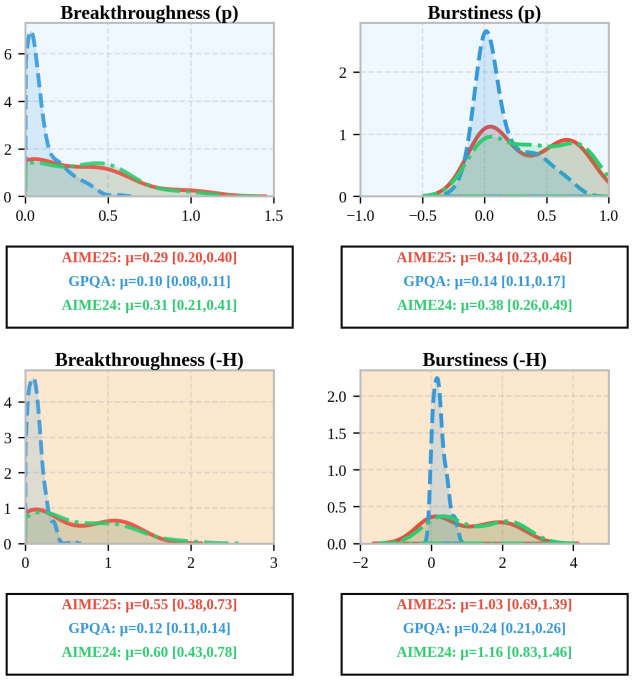
	AIME24 vs AIME25 ($n_1 = 30, n_2 = 30$)		AIME24 vs GPQA ($n_1 = 30, n_2 = 198$)		AIME25 vs GPQA ($n_1 = 30, n_2 = 198$)	
Emergence Score	U	p	U	p	U	p
Breakthroughness (p)	501	0.455	4523	4.0×10^{-6}	3471	0.137
Burstiness (p)	516	0.333	4191	5.4×10^{-6}	3568	0.010
Breakthroughness (-H)	501	0.455	5581	8.9×10^{-15}	4556	2.5×10^{-6}
Burstiness (-H)	492	0.540	5104	9.2×10^{-14}	3883	3.6×10^{-4}

Table 2

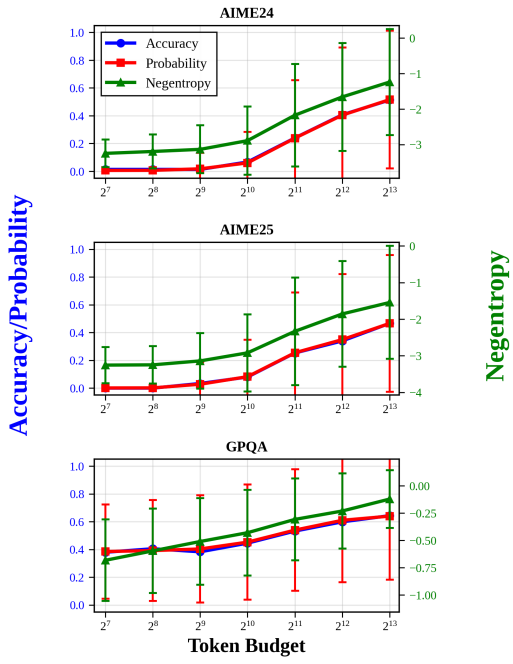
C.2. Analysis for Phi-4-Reasoning-Plus and QwQ-32B

This subsection presents the analysis of emergent inference-time scaling across datasets for Phi-4-Reasoning-Plus and QwQ-32B, following the same methodology as presented for Deepseek-R1-Distill-Qwen-32B in Section C.1. Both models are evaluated across GPQA, AIME24, and AIME25, with emergence score distributions and aggregate scaling curves shown in Figures 11 and 12. Mann-Whitney U tests for statistical significance are presented in Tables 3 and 4.

C.2.1. Phi-4-Reasoning-Plus



(a) Emergence score distributions for Phi-4-Reasoning-Plus across all evaluated datasets.



(b) Aggregate scaling curves for Phi-4-Reasoning-Plus.

Figure 11. Emergence analysis for Phi-4-Reasoning-Plus across all evaluated datasets, showing (a) distributions of emergence scores and (b) aggregate scaling behaviour for the same datasets. The vertical line separates the two panels for visual clarity.

The results for Phi-4-Reasoning-Plus show patterns consistent with those observed for Deepseek-R1-Distill-Qwen-32B in the main text. Visual inspection of Section C.2.1 reveals that AIME24 and AIME25 exhibit significantly higher emergence

	AIME24 vs AIME25 ($n_1 = 30, n_2 = 30$)		AIME24 vs GPQA ($n_1 = 30, n_2 = 198$)		AIME25 vs GPQA ($n_1 = 30, n_2 = 198$)	
Emergence Score	U	p	U	p	U	p
Breakthroughness (p)	477	0.695	3940	0.004	3663	0.040
Burstiness (p)	499	0.473	4001	3.0×10^{-4}	3788	0.003
Breakthroughness (-H)	489	0.569	4795	6.0×10^{-8}	4280	1.0×10^{-4}
Burstiness (-H)	503	0.438	4395	1.3×10^{-6}	3697	0.008

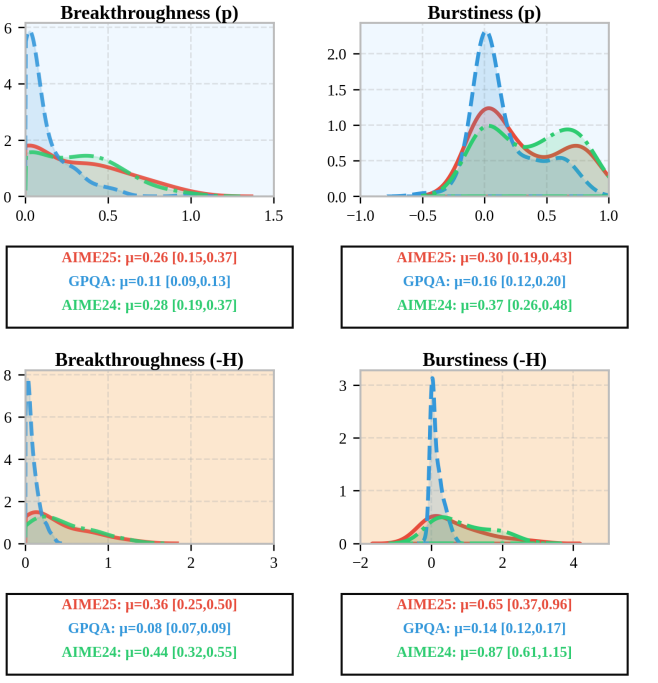
Table 3. Mann-Whitney U tests for statistical significance of emergence score distributions across datasets for Phi-4-Reasoning-Plus. Green indicates statistically significant differences ($p < 0.05$), red indicates non-significant differences ($p \geq 0.05$).

scores than GPQA across all metric-emergence score pairings, with distributions for the AIME datasets concentrated at higher values.

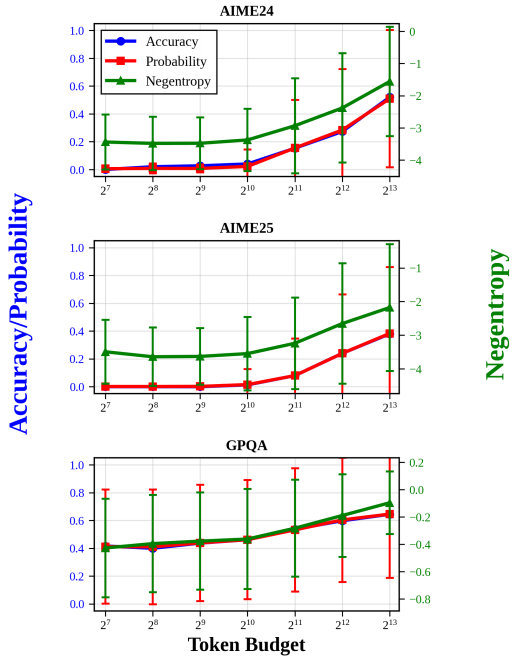
Statistical testing confirms these observations. Mann-Whitney U tests (Table 3) reveal significant differences between AIME24 and GPQA for all four emergence scores (Breakthroughness and Burstiness for both ground truth probability and negative entropy, with $p < 0.05$ in all cases). Similarly, AIME25 shows significant differences from GPQA across all emergence scores. However, no significant difference is found between AIME24 and AIME25 for any emergence score ($p > 0.4$ for all comparisons), suggesting these two datasets exhibit similar levels of emergent inference-time scaling behaviour.

The aggregate scaling curves in Section C.2.1 corroborate these findings, with GPQA showing smooth, gradual returns to additional inference-time compute, while AIME24 and AIME25 exhibit sharper increases in performance metrics at critical token budgets.

C.2.2. QwQ-32B



(a) Emergence score distributions for QwQ-32B across all evaluated datasets.



(b) Aggregate scaling curves for QwQ-32B.

Figure 12. Emergence analysis for QwQ-32B across all evaluated datasets, showing (a) distributions of emergence scores and (b) aggregate scaling behaviour for the same datasets. The vertical line separates the two panels for visual clarity.

	AIME24 vs AIME25 ($n_1 = 30, n_2 = 30$)		AIME24 vs GPQA ($n_1 = 30, n_2 = 198$)		AIME25 vs GPQA ($n_1 = 30, n_2 = 198$)	
Emergence Score	U	p	U	p	U	p
Breakthroughness (p)	489	0.569	3915	0.005	3499	0.116
Burstiness (p)	508	0.395	3592	0.002	3267	0.038
Breakthroughness (-H)	536	0.206	5203	3.3×10^{-11}	4588	1.5×10^{-6}
Burstiness (-H)	551	0.137	4199	1.6×10^{-7}	3216	0.049

Table 4. Mann-Whitney U tests for statistical significance of emergence score distributions across datasets for QwQ-32B. Green indicates statistically significant differences ($p < 0.05$), red indicates non-significant differences ($p \geq 0.05$).

QwQ-32B exhibits similar trends to both Deepseek-R1-Distill-Qwen-32B and Phi-4-Reasoning-Plus, though with some nuanced differences. As shown in Section C.2.2, emergence score distributions for AIME24 and AIME25 are again shifted toward higher values compared to GPQA, indicating more pronounced emergent scaling behaviour on the AIME datasets.

Mann-Whitney U tests (Table 4) reveal significant differences between AIME24 and GPQA for all emergence scores. For AIME25 versus GPQA, the pattern is slightly more nuanced: while Breakthroughness (p) does not reach statistical significance ($p = 0.116$), all other emergence scores show significant differences. Consistent with findings for the other models, no significant differences are observed between AIME24 and AIME25 for any emergence score, suggesting these datasets elicit comparable emergent scaling patterns from QwQ-32B.

The aggregate scaling curves in Section C.2.2 show that GPQA exhibits smoother scaling compared to the AIME datasets, which display sharper transitions in performance at critical token budgets, consistent with the higher emergence scores observed for these datasets.

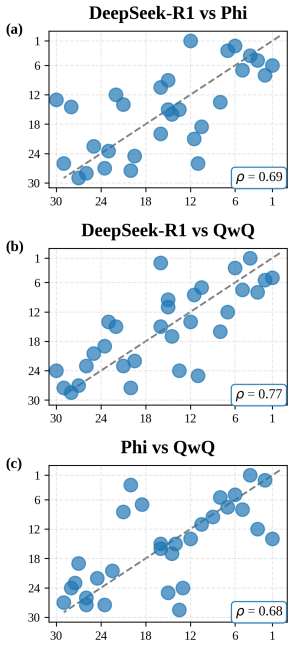
C.2.3. Summary

The analyses for Phi-4-Reasoning-Plus and QwQ-32B corroborate the findings presented for Deepseek-R1-Distill-Qwen-32B in the main text. Across all three models, AIME24 and AIME25 consistently exhibit more pronounced emergent inference-time scaling compared to GPQA, as evidenced by higher emergence scores and statistical tests. The two AIME datasets show similar emergence patterns to each other, suggesting that the observed differences in emergent scaling are primarily driven by dataset characteristics rather than model-specific behaviour. This cross-model consistency strengthens the conclusion that certain benchmark characteristics—potentially including compositional complexity or critical thinking budget requirements—are associated with more emergent inference-time scaling behaviour.

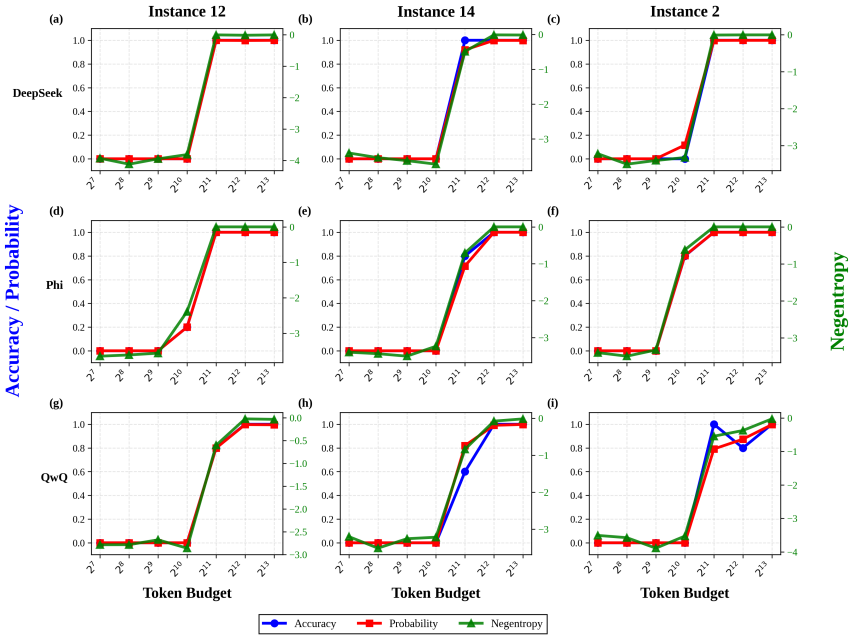
D. Instance Trends - Additional Analysis

This section presents additional instance-level analysis for AIME24 and GPQA, following the methodology described for AIME25 in Section 3.2.

D.1. AIME24 Instance Analysis



(a) Emergence score correlation.



(b) Top 3 AIME24 Instances.

Figure 13. Instance-level analysis for AIME24, showing (a) Spearman correlations of emergence score rankings between models, and (b) scaling curves for the top-3 instances exhibiting emergent inference-time scaling when evaluated with Deepseek-R1-Distill, across all models.

Section D.1 shows Spearman correlations for the emergence scores of AIME24 instances between different models. Consistent with the findings for AIME25 in the main text, strong correlation coefficients are observed across all model pairings (between $0.90 - 0.94$). This provides further evidence that the same problem instances consistently exhibit emergent inference-time scaling across different reasoning models, suggesting that emergence is primarily a property of individual problems rather than model-specific characteristics.

Section D.1 shows scaling curves for the top-3 AIME24 instances exhibiting emergent inference-time scaling when evaluated with Deepseek-R1-Distill, across all three models. Similar to AIME25, we observe sharp transitions in accuracy, ground truth probability, and negative entropy at critical token budgets. The timing of these transitions varies across models, with Phi-4-Reasoning-Plus generally showing the earliest transitions, followed by Deepseek-R1-Distill, and then QwQ-32B. The corresponding input prompts are shown in Table 5.

D.2. GPQA Instance Analysis

Section D.2 shows Spearman correlations for emergence scores of GPQA instances between models. In contrast to the AIME datasets, GPQA exhibits considerably weaker cross-model correlations. While Phi-4-Reasoning-Plus and QwQ-32B show moderate correlation ($0.65 - 0.71$), Deepseek-R1-Distill shows notably weak correlation with other models (≈ 0.18). This suggests that for GPQA, which instances exhibit emergent scaling behaviour is more model-dependent than for the AIME datasets.

This finding is consistent with the dataset-level analysis in Section 3.1, which showed that GPQA exhibits less pronounced emergent inference-time scaling overall. The weaker cross-model correlations may indicate that GPQA problems do

Instance Index	Problem
11	Every morning Aya goes for a 9-kilometer-long walk and stops at a coffee shop afterwards. When she walks at a constant speed of s kilometers per hour, the walk takes her 4 hours, including t minutes spent in the coffee shop. When she walks $s + 2$ kilometers per hour, the walk takes her 2 hours and 24 minutes, including t minutes spent in the coffee shop. Suppose Aya walks at $s + \frac{1}{2}$ kilometers per hour. Find the number of minutes the walk takes her, including the t minutes spent in the coffee shop.
14	Consider the paths of length 16 that follow the lines from the lower left corner to the upper right corner on an 8×8 grid. Find the number of such paths that change direction exactly four times, as in the examples shown below.
2	Jen enters a lottery by picking 4 distinct numbers from $S = \{1, 2, 3, \cdots, 9, 10\}$. 4 numbers are randomly chosen from S. She wins a prize if at least two of her numbers were 2 of the randomly chosen numbers, and wins the grand prize if all four of her numbers were the randomly chosen numbers. The probability of her winning the grand prize given that she won a prize is $\frac{m}{n}$ where m and n are relatively prime positive integers. Find $m + n$.

Table 5. AIME24 top-k instances displaying emergent inference-time scaling

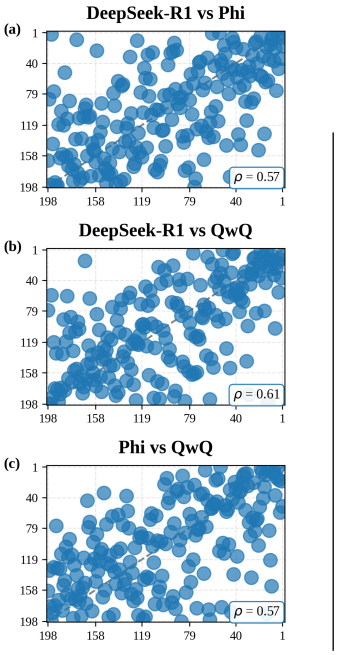
not possess the same structural characteristics that consistently elicit emergent scaling across different reasoning models. Alternatively, the lower overall emergence scores for GPQA may simply provide insufficient signal to detect consistent cross-model patterns.

Section D.2 shows scaling curves for the top-3 GPQA instances exhibiting emergent inference-time scaling when evaluated with Deepseek-R1-Distill. Compared to the AIME datasets, the transitions are generally less sharp and more variable across models, consistent with the weaker cross-model correlations observed. The corresponding input prompts are shown in Table 6.

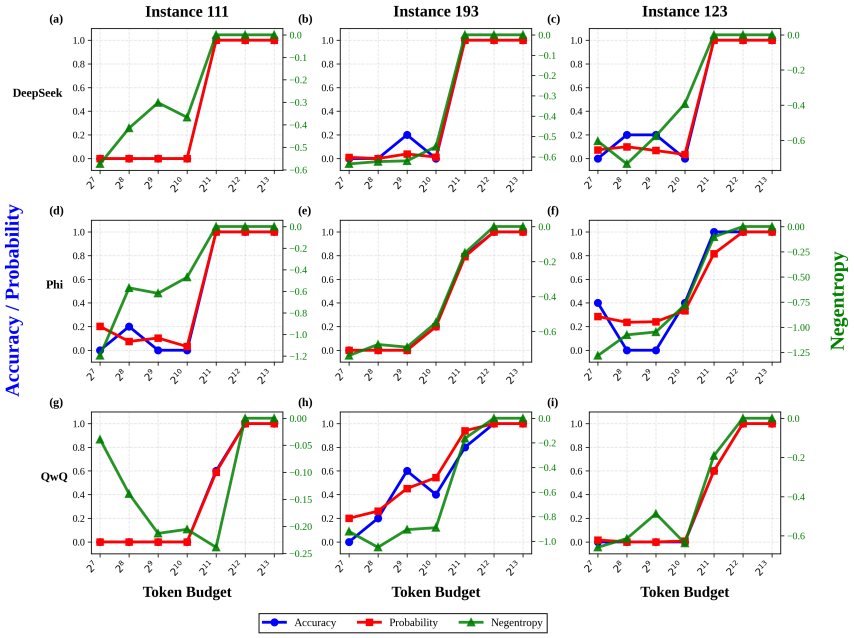
D.3. Summary

The instance-level analysis reveals important differences between datasets. For AIME24 and AIME25, strong cross-model correlations ($0.88 - 0.94$) indicate that specific problem instances consistently exhibit emergent inference-time scaling regardless of the model used. This suggests that emergence is primarily determined by problem characteristics rather than model-specific properties.

In contrast, GPQA shows weaker and more variable cross-model correlations, particularly involving Deepseek-R1-Distill. This may reflect either: (1) fundamental differences in how different models approach GPQA problems, or (2) the overall lower prevalence of emergent scaling on GPQA providing insufficient signal to detect consistent patterns. These findings complement the dataset-level analysis and suggest that the compositional, multi-step nature of AIME problems may be particularly conducive to consistent emergent scaling behaviour across models.



(a) Emergence score correlation.



(b) Top 3 GPQA Instances.

Figure 14. Instance-level analysis for GPQA, showing (a) Spearman correlations of emergence score rankings between models, and (b) scaling curves for the top-3 instances exhibiting emergent inference-time scaling when evaluated with Deepseek-R1-Distill, across all models.

Instance Index	Problem
111	Let $\|\alpha\rangle$ be the state describing an electron, such that it is proportional to $(1+i)\|up\rangle + (2-i)\|down\rangle$, where $\|up\rangle$ and $\|down\rangle$ are the eigenstates of the z-projection of the spin operator. Calculate the probability of measuring the particle in each of the eigenstates of the operator whose matrix representation is given by the elements A_{ij}, such that $A_{ij} = \frac{\hbar}{2}$ if $i \neq j$, and 0 otherwise. Also, find the average value of that operator. Choose from one of the four options below. A) $0.61, 0.29$ and $\frac{2\hbar}{\sqrt{7}}$ B) $0.28, 0.72$ and $\frac{\hbar}{\sqrt{7}}$ C) $0.54, 0.46$ and $\frac{3\hbar}{\sqrt{7}}$ D) $0.64, 0.36$ and $\frac{\hbar}{7}$
193	Consider a system of three spins S_1, S_2 and S_3, each of which can take spin $+1$ and -1. The energy of the system is given by $E = -J[S_1S_2 + S_1S_3 + S_2S_3]$. Find the partition function Z of the system. ($\beta = 1/kT$, k = Boltzmann constant and T = temperature) Choose from one of the four options below. A) $Z = 6e^{2J\beta} + 2e^{-2J\beta}$ B) $Z = 2e^{2J\beta} + 6e^{-2J\beta}$ C) $Z = 2e^{3J\beta} + 6e^{-J\beta}$ D) $Z = 2e^{-3J\beta} + 6e^{J\beta}$
123	Particles are collided at the center of a spherical detector, producing a new type of particles that travel uninterrupted at ultra-relativistic velocities highly centered around the Lorentz factor of $\gamma \sim 20$. On average, one-third of these fast-decaying particles reach the detector inner walls before they decay. The radius of the detector is 30 meters. What Lorentz factor is needed in order to have about two-thirds of these particles reach the detector inner walls? Choose from one of the four options below. A) 54 B) 40 C) 68 D) 28

Table 6. GPQA top-k instances displaying emergent inference-time scaling

E. Scale Trends - Additional Analysis

E.1. Statistical Tests for Model Size Distributions - AIME25

This appendix presents further statistical analysis of the effect of model size on emergent inference-time scaling.

We first present the results of Mann-Whitney U tests for statistical significance between model size pairings. These are shown for each metric-emergence score pairing in the tables below.

Table 7. Mann-Whitney U test results across model scales. Green: significant ($p < 0.05$), Red: non-significant ($p \geq 0.05$).

(a) Breakthroughness (p)

	1.5B	7B	14B
7B	$p = 0.241$ $U = 498$	—	—
14B	$p = 0.075$ $U = 548$	$p = 0.214$ $U = 504$	—
32B	$p = 0.045$ $U = 565$	$p = 0.214$ $U = 504$	$p = 0.562$ $U = 440$

(b) Burstiness (p)

	1.5B	7B	14B
7B	$p = 0.120$ $U = 530$	—	—
14B	$p = 0.100$ $U = 537$	$p = 0.295$ $U = 487$	—
32B	$p = 0.001$ $U = 656$	$p = 0.039$ $U = 570$	$p = 0.117$ $U = 531$

(c) Breakthroughness (-H)

	1.5B	7B	14B
7B	$p = 0.479$ $U = 454$	—	—
14B	$p = 0.013$ $U = 601$	$p = 0.056$ $U = 558$	—
32B	$p = 0.194$ $U = 509$	$p = 0.321$ $U = 482$	$p = 0.806$ $U = 392$

(d) Burstiness (-H)

	1.5B	7B	14B
7B	$p = 0.544$ $U = 443$	—	—
14B	$p = 0.003$ $U = 633$	$p = 0.010$ $U = 608$	—
32B	$p = 0.103$ $U = 536$	$p = 0.117$ $U = 531$	$p = 0.834$ $U = 385$

The results of the Mann-Whitney U tests on the ground truth probability metric find fewer significant differences than implied by the figures in Section 3.3. However, the failure of some of these tests to reach significance should be interpreted with caution. Due to the non-parametric nature of Mann-Whitney U tests, and with only 30 samples per model size, this analysis may not have sufficient statistical power to detect small-to-medium effect sizes, particularly when distributions have high variance or substantial overlap. The latter is likely for adjacent model sizes (e.g, 7B v.s. 14B) where effects may exist but are subtle. Therefore, we compute Cohen's d for each model size pairing in Table 8 below. Cohen's d is a measure of effect size independent of sample size, thus correcting for some of the issues of the Mann-Whitney U tests.

Table 8. Cohen's d effect sizes across model scales. Red: negligible ($d < 0.2$), Yellow: small ($0.2 \leq d < 0.5$), Green: medium-to-large ($d \geq 0.5$).

(a) Breakthroughness (p)

	1.5B	7B	14B
7B	$d = 0.298$	—	—
14B	$d = 0.561$	$d = 0.215$	—
32B	$d = 0.673$	$d = 0.330$	$d = 0.127$

(b) Burstiness (p)

	1.5B	7B	14B
7B	$d = 0.384$	—	—
14B	$d = 0.662$	$d = 0.269$	—
32B	$d = 0.813$	$d = 0.405$	$d = 0.131$

(c) Breakthroughness (-H)

	1.5B	7B	14B
7B	$d = 0.192$	—	—
14B	$d = 0.692$	$d = 0.405$	—
32B	$d = 0.423$	$d = 0.224$	$d = -0.132$

(d) Burstiness (-H)

	1.5B	7B	14B
7B	$d = 0.121$	—	—
14B	$d = 0.830$	$d = 0.587$	—
32B	$d = 0.487$	$d = 0.329$	$d = -0.201$

Cohen's d results align with the results of Section 3.3; we see large differences in ground truth probability emergence distributions when comparing the 1.5B model to the 14B and 32B models, and a more subtle difference between the 7B and 14B/32B models; with a negligible difference between the 14B and 32B models.

E.2. Analysis for AIME24

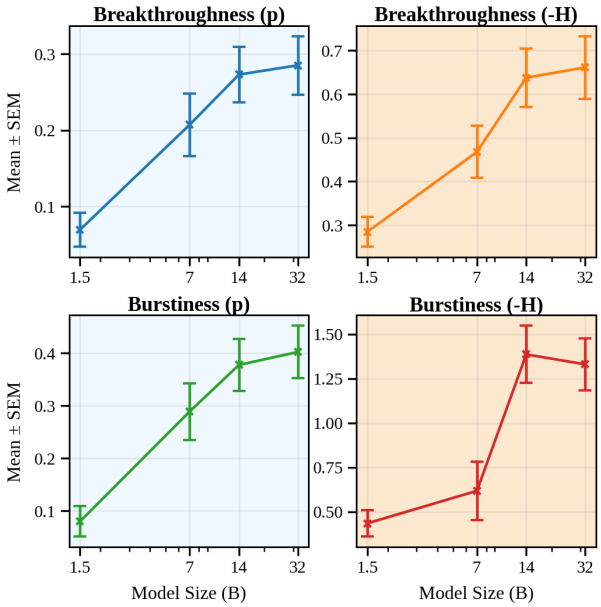


Figure 15. Scaling curve for AIME24.

Table 9. Mann-Whitney U test results across model scales for AIME24. Green: significant ($p < 0.05$), Red: non-significant ($p \geq 0.05$).

(a) Breakthroughness (p)

	1.5B	7B	14B
7B	$p = 0.007$ $U = 617$	—	—
14B	$p < 0.001$ $U = 713$	$p = 0.126$ $U = 528$	—
32B	$p < 0.001$ $U = 683$	$p = 0.163$ $U = 517$	$p = 0.497$ $U = 451$

(b) Burstiness (p)

	1.5B	7B	14B
7B	$p = 0.002$ $U = 651$	—	—
14B	$p < 0.001$ $U = 730$	$p = 0.149$ $U = 521$	—
32B	$p < 0.001$ $U = 741$	$p = 0.095$ $U = 539$	$p = 0.290$ $U = 488$

(c) Breakthroughness (-H)

	1.5B	7B	14B
7B	$p = 0.013$ $U = 602$	—	—
14B	$p < 0.001$ $U = 724$	$p = 0.025$ $U = 583$	—
32B	$p < 0.001$ $U = 724$	$p = 0.016$ $U = 596$	$p = 0.438$ $U = 461$

(d) Burstiness (-H)

	1.5B	7B	14B
7B	$p = 0.509$ $U = 449$	—	—
14B	$p < 0.001$ $U = 768$	$p = 0.001$ $U = 671$	—
32B	$p < 0.001$ $U = 762$	$p = 0.001$ $U = 659$	$p = 0.562$ $U = 440$

Table 10. Cohen's d effect sizes across model scales for AIME24. Red: negligible ($d < 0.2$), Yellow: small ($0.2 \leq d < 0.5$), Green: medium-to-large ($d \geq 0.5$).

(a) Breakthroughness (p)

	1.5B	7B	14B
7B	$d = 0.744$	—	—
14B	$d = 1.209$	$d = 0.304$	—
32B	$d = 1.233$	$d = 0.352$	$d = 0.058$

(b) Burstiness (p)

	1.5B	7B	14B
7B	$d = 0.869$	—	—
14B	$d = 1.321$	$d = 0.308$	—
32B	$d = 1.422$	$d = 0.392$	$d = 0.089$

(c) Breakthroughness (-H)

	1.5B	7B	14B
7B	$d = 0.678$	—	—
14B	$d = 1.191$	$d = 0.482$	—
32B	$d = 1.202$	$d = 0.527$	$d = 0.061$

(d) Burstiness (-H)

	1.5B	7B	14B
7B	$d = 0.259$	—	—
14B	$d = 1.364$	$d = 0.852$	—
32B	$d = 1.387$	$d = 0.826$	$d = -0.065$

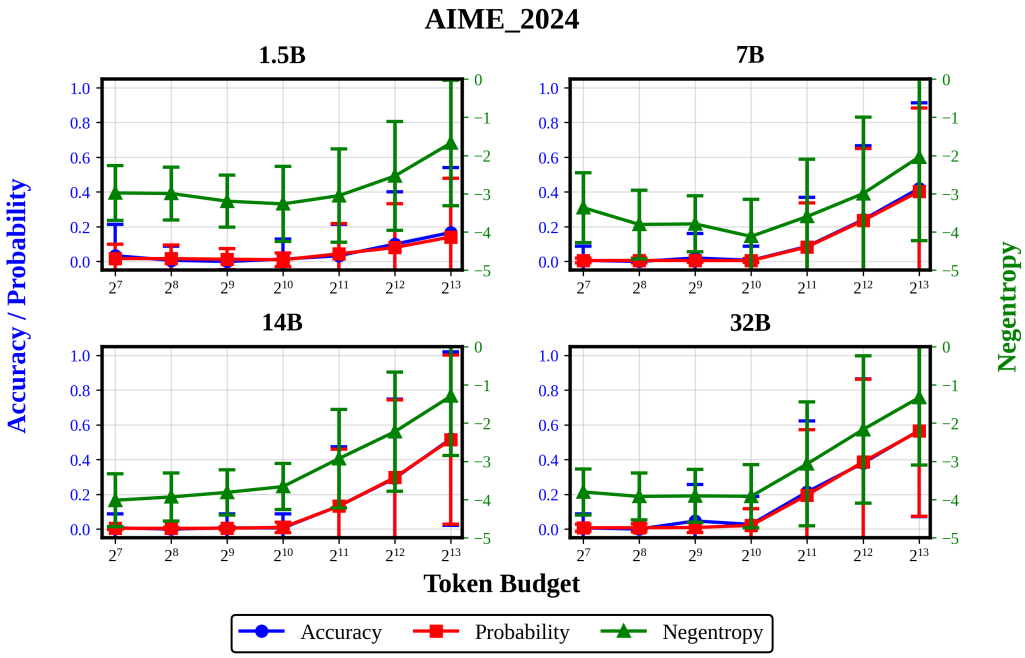


Figure 16. Aggregate scaling curves across multiple scales for AIME24.

F. Limitations

Limitations of this study are given below. Broadening the study to address these limitations and seeing whether similar emergent scaling behaviour occurs provides a natural direction for future research, alongside the suggestions of Section 4.

Only sequential scaling is used to increase inference-time compute budget: There are many ways to utilise compute at inference-time, some of which are covered in Section A. This study scaled inference-time compute *sequentially*, by appending special tokens to force reasoning traces to desired lengths. This design choice was made as it appears to be the most natural setting in which models may make sudden breakthroughs as they are reasoning, similar to "Aha!" moments exhibited by humans (Kounios & Beeman, 2009). However, it is plausible that similar breakthroughs could also occur with parallel scaling methods—for example, is there a critical branching factor when performing Monte Carlo Tree Search (MCTS) (Sutton et al., 1999), above which a model will converge to the correct solution, and below which it will not? These possibilities seem less likely to us compared to the sequential scaling method of having a model critique and revise its own answer, but they deserve further investigation given the popularity of parallel scaling methods and the positive results for sequential scaling observed in this study. The introduction of reward models to direct the search amongst solutions would add another layer of complexity to this analysis, as it adds additional hyperparameters that could exhibit thresholding behaviour.

Inverse scaling effects are not considered: We limit our study to reasoning traces of up to 8192 tokens due to computational constraints. However, it is likely that there are additional interesting scaling behaviours to be observed at higher budgets. For example, whilst the three main models used in this study (Deepseek-R1-Distill-32B, QwQ-32B, Phi-4-Reasoning-Plus) show signs of saturation on GPQA at the inference-time budgets of 2^{12} and 2^{13} tokens, this is not the case for AIME24 and AIME25, where even at 2^{13} tokens, the models maintain constant returns to additional inference-time compute (after having passed the critical token budget 2^{10} tokens). One key scaling trend that would set in at higher budgets is inverse scaling effects (Gema et al., 2025), where models exhibit negative returns to additional inference-time compute as they get stuck in repetitive loops, fail to carry out long chains of deductive reasoning, and get distracted by irrelevant information that is generated in the reasoning trace. As inverse scaling effects set in, it is likely that some datasets or dataset instances will see *sudden drops* in performance, which would motivate a reverse study of the present one to investigate when and why this happens.

Only mathematics and science benchmarks are evaluated: Our study focuses on mathematics and science benchmarks, specifically AIME24, AIME25, and GPQA-Diamond. These benchmarks were chosen as (a) they contain tasks that reasoning models are most proficient at (logical and multi-step reasoning), and (b) they have finite solution sets, allowing for the calculation of ground truth probability and negative entropy metrics normalised over the solution set. However, they represent a small fraction of the full spectrum of tasks that reasoning models are capable of solving. It is likely that emergent inference-time scaling occurs for a broader range of tasks than those evaluated in this study.