EMERGENT ABILITIES IN LARGE LANGUAGE MODELS AT INFERENCE TIME

Iyngkarran Kumar
University of Edinburgh
iyngkarrankumar@gmail.com

Edoardo Ponti
University of Edinburgh
eponti@ed.ac.uk

ABSTRACT

Scaling inference-time compute (also known as test-time compute) is an effective method to improve the performance of large language models. Allowing models to produce long chains of reasoning unlocks a "System 2 "-style approach to problems, which has dramatically improved model performance on a wide range of benchmarks, particularly mathematics, scientific, and coding tasks. Existing work has established scaling laws for inference compute, but has not investigated cases of sharp, abrupt improvements in model performance as inference compute is scaled. These cases, which have previously been called "emergent abilities" when studied under training compute scaling, are the focus of this study. We study the frequency of emergent test-time scaling across datasets (AIME24, AIME25, GPQA), model families (Deepseek-R1-Distill, QwQ, Phi-4-Reasoning-Plus), and model sizes (1.5B, 7B, 14B, 32B), as well as individual dataset instances. We find substantial variation in emergent scaling across datasets—with AIME24 and AIME25 exhibiting significantly sharper scaling behaviour than GPQA at the aggregate level. This suggests that in certain cases, piecewise linear curves better model the relationship between accuracy and $\log(compute)$ than simple linear fits suggested in previous work. When evaluating individual problems from the datasets, we find that the same instances consistently exhibit sharp, non-linear scaling, indicating that the degree of emergent test-time scaling observed is dependent on features of individual problems. We also find that there is no significant variation in the abruptness or magnitude of emergent scaling across model families, but the compute budget at which this abrupt improvement occurs differs. Finally, we find that larger models are more likely to see emergent scaling of accuracy than smaller models. Our results lay the groundwork for understanding when unpredictable breakthroughs occur in language models as test-time compute is increased, and contribute to a broader understanding of the predictability and continuity of test-time scaling.

1 Introduction

Inference-time, or test-time, scaling is an established method to improve the performance of large language models. By allowing a model to produce long reasoning traces before reporting a final answer, sampling multiple responses first before aggregating them, or using a combination of these two methods, 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.

The effectiveness of scaling test-time compute to improve language model performance continues the trend of scaling key model inputs to achieve qualitatively new abilities. Previously, this has come in the training domain, with existing research establishing mathematical relationships ("scaling laws") that link model size, training dataset size, and training

compute to pre-training loss (Kaplan et al., 2020; Hoffmann et al., 2022). The training scaling laws have proven to be robust over multiple orders of magnitude of training inputs, allowing some model abilities to be predicted by training counterparts with $1,000 - 10,000\times$ less compute (Achiam et al., 2023). Reliable scaling behaviour has scientific, economic, and safety benefits, enabling developers to anticipate computational requirements, as well as model abilities and potential risks without needing to commit significant resources for training.

However, a number of studies have identified "emergent abilities" in large language models when scaling up training inputs—specific tasks for which models exhibit sharp, abrupt improvements when scaled past a certain threshold (Wei et al., 2022a; Srivastava et al., 2022; Ganguli et al., 2022; Berti et al., 2025). For example, when studying the performance of the Chinchilla and Gopher (Hoffmann et al., 2022; Rae et al., 2021) families on a suite of MMLU tasks (Hendrycks et al., 2020), Wei et al. (2022a) find that model performance remains constant at random choice for model sizes smaller than 10B parameters, but sudden improvement occurs when scaled to 100B parameters. The existence of emergent abilities at train-time poses scientific questions—why do some tasks exhibit sharp increases in performance in contrast to smooth scaling?—as well as engineering and safety problems; ideally, models would exhibit reliable and predictable scaling across all tasks and training inputs, with minimal surprises to both developers and end users when scaling up. However, the nature of emergent abilities has been questioned with some arguing that they are an artifact of discrete evaluation metrics, rather than a feature of tasks and models themselves Schaeffer et al. (2023).

These questions and concerns are relevant in the test-time setting too. Existing work has posited that the relationship between test-time compute budget and model performance is approximately log-linear (Snell et al., 2024; Muennighoff et al., 2025; Brown et al., 2025), but has carried out these scaling studies on a limited number of benchmarks, and aggregated scaling behaviour across a wide number of tasks, which may smooth out sharp scaling behaviour that occurs at a more granular level. This study is thus concerned with sharp, abrupt increases in model performance as test-time compute budget is scaled. The existence of such scaling behaviour would challenge the assumption that test-time compute scaling is well modelled with a linear relationship between the logarithm of test-time compute budget and model performance—a point that has already been frequently made with respect to train-time scaling (Caballero et al., 2023; Alabdulmohsin et al., 2022; Villalobos, 2023). Additionally, it would raise questions about sudden breakthroughs in reasoning traces—"Aha!" moments present in the model's computation that resemble sudden moments of insight in human cognition (Kounios and Beeman, 2009). This study aims to identify the key factors that determine whether test-time scaling is either smooth, continuous, and predictable, or sharp, non-linear, and abrupt[1].

1.1 Research questions addressed in this study

More precisely, the research questions addressed in this study are:

1. To what extent does emergent scaling vary across different datasets? Are the returns to test-time compute best modelled by a log-linear relationship, or are there cases where other fits (such as piecewise linear) are more appropriate?
2. To what extent do individual dataset instances consistently exhibit emergent test-time scaling? Across multiple models, do the same instances consistently exhibit sharp, abrupt increases in performance?
3. To what extent does emergent scaling vary across different model families? Do some reasoning models display sharper scaling behaviour than others, or is this behaviour consistent across model families?
4. To what extent does emergent scaling vary across model size (number of parameters)? Do larger models exhibit sharper scaling than smaller ones, or does model size have a negligible influence on emergent test-time scaling?
5. Is the presence of emergent test-time scaling contingent on the metric used for evaluation? Does emergent scaling only occur when using discrete evaluation metrics, or are they present when tracking continuous evaluation metrics too?

[1]Code for this project is available at `https://github.com/IyngkarranKumar/test_time_emergent_scaling`

1.2 Structure of this study

Section 2 first introduces key background concepts, such as existing definitions of emergent scaling, proposed explanations for emergent scaling, as well as the current literature on methods to use compute at test-time, and test-time scaling laws. Section 3 describes our approach to studying emergent abilities at test-time. We scale test-time compute sequentially—that is, allowing models to produce longer reasoning traces—by appending special tokens onto the end of a reasoning trace to cut short or extend the reasoning (Section 3.1). We evaluate three popular open-weight reasoning models—Deepseek-R1-Distill-32B, QwQ-32B, and Phi-4-Reasoning-Plus (14B)—on mathematics and science benchmarks—AIME24, AIME25, and GPQA (Section 3.2). We track three metrics across token budget—accuracy, ground truth probability, and negative entropy over closed solution sets—the first of which is discrete, and the second and third are continuous measures of model performance. We then score the degree of emergent test-time scaling using two metrics: Breakthroughness$^+$ and Weighted Difference Skewness (Section 3.4), which capture two key aspects of emergent scaling behaviour. Section 4 presents the main empirical findings of the study, which are briefly summarised below. Section 5 discusses the implications of the results, this study's limitations, and avenues for future work. Finally, Section 6 concludes.

1.3 Contributions of this study

To our knowledge, this study is the first that directly investigates the presence of sharp, abrupt increases in model performance with respect to test-time compute. Our contributions are as follows:

- We introduce a new metric for scoring emergent scaling behaviour, Breakthroughness$^+$, that builds upon previous work in Srivastava et al. (2022) (Section 3.4). This metric captures the ratio of the overall change in a response variable to the average change between consecutive points, but is more robust to different types of scaling behaviour than the previous formulation.

- We find that emergent scaling varies substantially across datasets, with AIME24 and AIME25 displaying an abrupt increase in accuracy after 2^{10} thinking tokens, whilst GPQA exhibits much smoother scaling behaviour. This suggests that, in some cases, two-piece piecewise linear curves are better fits to the accuracy-compute relationship than simple linear fits previously suggested (Snell et al., 2024; Muennighoff et al., 2025; Wu et al., 2025) (Section 4.1).

- We find a strong correlation across models between dataset instances that display the sharpest scaling behaviour. This suggests that features of individual problems, rather than dataset-level characteristics, have a strong influence on the degree of emergent test-time scaling (Section 4.2).

- We find that emergent scaling trends are mostly consistent across model families, with the abruptness and magnitude of scaling behaviour similar across Deepseek-R1, QwQ, and Phi-4-Reasoning-Plus. However, the compute budget at which sharp scaling occurs varies across models, with QwQ seeing a later onset than Deepseek and Phi (Section 4.3). This suggests that differences between model families, such as architecture and training data, do not have a strong influence on the "global" emergent scaling trends—namely the abruptness and magnitude of the increase in performance—but do exert an influence on the token budget at which this occurs (Section 4.3).

- We find that larger models show sharper scaling of accuracy and ground truth probability than smaller models, suggesting that the test-time scaling of larger models is less predictable than smaller models (Section 4.4).

- We find sharp, abrupt increases in both discrete and continuous metrics, namely accuracy (discrete), ground truth probability (continuous), and negative entropy (continuous). This suggests that emergent scaling behaviour is not an artefact of the metric choice, but a feature of the models and tasks themselves.

2 Background

This section reviews the existing literature on test-time compute scaling (Section 2.1), and emergent abilities in large language models (Section 2.2).

2.1 Scaling test-time compute in language reasoning models

2.1.1 Using compute at test-time

There is a wide range of methods for leveraging compute at test-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). Self-critique and deliberation amongst solutions allow a model to move past "System 1" methods of immediate pattern matching towards a "System 2 "-style reasoning process (Zhang et al., 2025).

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). The quality of a parallel scaling method can be evaluated with respect to its Coverage—the chance that a correct solution is generated when sampling multiple responses, and Aggregation Quality—the ability to identify a correct solution amongst a large number of sampled responses. 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 test-time can enhance a model's performance across numerous domains, but well-structured tasks that require multi-step reasoning and verification particularly benefit. This includes mathematical tasks which have substantial logical reasoning requirements—such as MATH-500 (Zhang et al., 2024) and OlympiadBench (He et al., 2024), programming and code generation (Jimenez et al., 2023; Jain et al., 2024), strategic game playing (Costarelli et al., 2024), and scientific reasoning (Rein et al., 2024; Feng et al., 2025).

2.1.2 Scaling laws for test-time compute

The proliferation of methods to leverage test-time compute has motivated work that seeks to quantitatively understand the relationship between scaling test-time compute and model performance. Multiple studies posit an approximately linear relationship between the logarithm of test-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, with the model entering into repetitive loops, getting distracted by irrelevant information, and failing to maintain consistency when carrying out long chains of deductive reasoning—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, and notably, there can be minimal degradation in model performance when reducing the trace length by up to 50%.

Given the wide range of methods to scale test-time compute, studies such as Snell et al. (2024) and Wu et al. (2025) have studied how to optimally allocate compute at test-time. Snell et al. (2024) studies the optimal allocation of test-time compute amongst various inference strategies for Palm 2 (Anil et al., 2023) on the MATH (Hendrycks et al., 2021), 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 (i.e: 4x less compute to achieve the same level of performance). Wu et al. (2025) broadens this to more models and benchmarks, and introduces Reward Balanced

Search (REBASE)—a tree search method that explores nodes based on a process-based reward model, which is compute-optimal relative to standard sampling and MCTS methods.

To situate the current study within the literature, a growing body of work has looked to quantify the relationship between using additional compute at inference-time and model performance. However, to our knowledge, there are no studies that explicitly focus on the degree of unpredictability that is present in test-time scaling of language reasoning models (LRMs)—specifically, the extent to which sharp, abrupt improvements in model performance occur with larger test-time compute budgets. Relatedly, many existing studies have aggregated scaling behaviour across entire/and or multiple datasets, which can mask the presence of emergent scaling. This work seeks to address these limitations. Additionally, we are interested in the nature of language model "reasoning" itself—specifically, whether these systems display sudden breakthroughs in their reasoning traces, akin to "Aha!" moments of insight displayed by humans (Kounios and Beeman, 2009).

We now review the existing literature on emergent abilities in large language models at train-time—where much previous work in this area has been focused.

2.2 Emergent abilities of large language models

2.2.1 Key definitions and phenomena

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.*" (emphasis our own). The definition of Wei et al. (2022a) is different from the one used in natural sciences more broadly, where a phenomenon is described as emergent if it cannot be predicted by analysing the microscale interactions between the constituents of a system, and is instead only observed at the macroscale (Anderson, 1972). The definition of emergent scaling used in this work, which is consistent across previous studies, is a sharp, abrupt improvements in model performance with respect to a scaling quantity.

Wei et al. (2022a); Ganguli et al. (2022) observe that a large number of tasks, mostly within the BIG-Bench (Srivastava et al., 2022) and MMLU (Hendrycks et al., 2020) benchmarks that exhibit emergent scaling behaviour with respect to model parameters and/or training compute, such as few-shot 3-digit addition and French-English translation. Additionally, the majority of these tasks were evaluated in the few-shot setting, or using other prompting strategies, such as instruction-tuning (Wei et al., 2021) or allowing the model access to a scratch pad (Nye et al., 2021). For example, Brown et al. (2020) shows that a 6B model achieves only 1% accuracy on few-shot 3-digit addition, a 13B model improves to 8%, but a 175B model makes a drastic improvement to 80%. For a comprehensive list of 137 abilities identified as emergent with respect to training inputs, see Wei (2022).

2.3 Relationship between emergent abilities, discrete, and continuous metrics

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—those that only reward a model when it produces the correct answer—with continuous metrics, which recognise and award partial progress towards a solution, emergent scaling largely disappears. However, a more nuanced exploration of the relationship between metric choice and emergent abilities is needed; follow-up work (Berti et al., 2025) has questioned the appropriateness of proposed continuous metrics in Schaeffer et al. (2023), such as Token Edit Distance, for evaluating model performance. Berti et al. (2025) also identified methodological issues

in the definition of emergent scaling used in Schaeffer et al. (2023), arguing that this work defines emergent scaling too strictly (Berti et al., 2025), ruling out many cases that would satisfy intuitive notions of emergence in the process. 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).

Schaeffer et al. (2023) also argue that emergent scaling can be partially attributed to evaluating on a test-set with a small number of samples; by simply increasing the test-set size, scaling curves become much smoother. This is experimentally studied in Hu et al. (2023), which samples $N = 10^5$ times from a model and measures pass@N, which results in smoother scaling curves. However, this method is limited by it's expensive computational requirements and the suitability of pass@N to robustly measure model performance.

2.4 Mechanisms underlying emergent scaling behaviour

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 and Goyal, 2023), 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 for natural language tasks in Arora and Goyal (2023), and experimentally studied in the vision (Okawa et al., 2024) and natural language (Lubana et al., 2024) domains. Okawa et al. (2024) study how a diffusion model learns how to independently generate shapes of distinct colour and shape, before quickly learning how to compose these concepts to generate out-of-distribution shapes. Lubana et al. (2024) show that sudden increases in performance on narrow natural language tasks can be attributed to a model learning "latent skills"—emergent scaling in language models—conceptualizing the learning of skills as a model's ability to connect two nodes on an abstract 'concept graph', leading to a proposed theory of emergent behavior based on graph percolation theory. Specifically, the percolation threshold p_c (the probability p_c of a randomly chosen edge being present at which a macroscopic cluster on the graph forms) is likened to a model suddenly learning to compose concepts that were not seen together in the training data (compositional generalisation).

Alternative explanations for emergent scaling have been proposed. Du et al. (2025) study the relationship between emergent scaling and pretraining loss, finding that once pretraining loss dropped below a certain threshold, performance on downstream tasks exhibited sharp, non-linear jumps. As noted by Berti et al. (2025), this highlights that emergent abilities are not just a function of model size, but are also influenced by training dynamics. Huang et al. (2024) links emergent abilities with the related phenomenon of grokking (Power et al., 2022) and deep double descent (Nakkiran et al., 2021), hypothesising that all three scaling behaviours can be explained by analysing competition between memorisation and generalisation circuits in neural models, again, reinforcing the point that training dynamics have a strong influence on the likelihood of emergent scaling.

Another proposed explanation for sharp scaling behaviour in language models is given in Wu and Lo (2024), which suggests that emergent scaling behaviour results from aggregating scaling behaviour across task distributions of varying complexity. They categorise tasks into those that exhibit U-shaped and inverted U-shaped scaling, and show that simple aggregations of these functions can produce the sudden "breakthrough" in model performance that characterises emergent scaling. Their hypothesised explanation for emergent scaling is intuitive, but was only tested on multiple-choice tasks and remains to be extended to other settings.

In summary, many different underlying causes of emergent scaling in language models have been proposed, some of which have been listed above. None of the proposed mechanisms for emergent scaling that are discussed above are mutually exclusive, or broadly accepted as the primary way that emergent scaling occurs; thus, this topic is still an active area of research.

3 Method

This section details the methodology used in this study. Section 3.1 gives an overview of the method used to scale test-time compute, followed by Section 3.2, which covers the datasets and models used in the study. Section 3.3

then discusses the discrete and continuous metrics that are tracked over test-time compute budgets, before Section 3.4 introduces two metrics that are used to quantitatively score the extent to which a particular scaling curve exhibits emergent scaling behaviour.

3.1 Scaling test-time compute

This study focuses on sequential scaling of test-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 and Beeman, 2009).[2] To control the length of a reasoning trace, we use the budget forcing method of Muennighoff et al. (2025), appending special tokens onto the end of a reasoning trace to force a model to end, or continue, its generation. Specifically, 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 the `<EOS>` token, we remove the `<EOS>` token and append the sequence `Hmm, let me keep thinking` to force the model to reconsider it's answer and continue working. Once the model nears the token budget T, we append the sequence `The final answer is:` to force the model to output a solution, allowing the model $N_f = 10$ tokens to provide a solution[3], then end generation.

3.2 Datasets and models

We study scaling behaviour on math and science benchmarks, due to the proficiency of language reasoning models (LRMs) on these tasks (Xu et al., 2025). The benchmarks used in this study are:

1. American Invitational Mathematics Examination (AIME) 2024 and 2025 (Zhang and Math-AI, 2024, 2025) . These datasets contain 30 challenging mathematical problems with integer solutions in the range 0-999.
2. GPQA-Diamond—GPQA (Rein et al., 2024) is a multiple-choice benchmark of 448 questions testing domain expertise in biology, physics, and chemistry. The diamond split is a more challenging subset of the benchmark.

These datasets are also 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 test-time compute budgets (discussed in the next section).
We study test-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 study emergent test-time scaling 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).

3.3 Metrics

As noted in the previous section, the benchmarks in this study have closed solution sets: AIME-24 and AIME-25 have ground truths $y \in \{0, 1, 2, \ldots, 999\}$ and GPQA has $y \in \{A, B, C, D\}$, meaning that the model's output distribution over it's vocabulary can be renormalised over the solution set, denoted as $\mathcal{S}$. Letting M denote a reasoning model that

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

[3] As discussed in the next section, the candidate solutions that the model should output are either integers or multiple choice answers, which are typically between 1-3 tokens.

outputs a reasoning trace r and answer y^* to input prompt x, the model output is evaluated over the following **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 $\tilde{p}(s) = \frac{p_M(s|x,r)}{\sum_{s'\in\mathcal{S}} p_M(s'|x,r)}$ is the renormalised probability of solution s.

The accuracy metric is discrete, whereas the probability of ground truth and negative entropy metrics are continuous; all three metrics are tracked to determine if emergent scaling behaviour is observed for discrete metrics only, as suggested in Schaeffer et al. (2023), or appears across a wider range of measures. It is important to note that negative entropy is a measure of a model's confidence over the solution set, rather than a direct measure of model performance. This means that a model can have a high negative entropy score but a low accuracy/ground truth probability, if it is confident in the incorrect answer—a trend that sometimes occurs in our results (Section 4).

3.4 Scoring Emergent Scaling

This study uses two methods to score the degree to which a model exhibits sharp, abrupt (emergent) scaling behaviour. The first, Breakthroughness+, is an improved version of the Breakthroughness metric proposed in Srivastava et al. (2022), and captures the ratio of the total change in a response variable y to the average change between consecutive points. The second, Weighted Difference Skewness, is a measure of the symmetry of the distribution of metric differences ($\Delta y_i = y_{i+1} - y_i$). We now discuss in further detail the construction of each metric.

3.4.1 Scoring Emergent Scaling with 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:

$$B = \frac{s\cdot(\max_i y_i - \min_i y_i)}{\text{RootMedianSquare}(\{\Delta y_i\}_i)} \tag{1}$$

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 (2). This operator dampens rather than eliminates the influence of outlier differences, acting as a more practical averaging method than the root median square.

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

We make one further change to the Breakthroughness operator above. 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 RootMedian-Square 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:

$$B^+ = \frac{(\max_i y_i - \min_i y_i)^2}{\text{MeanSquareRoot}(\{\Delta y_i\}_i)} \tag{3}$$

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. We use the Weighted Difference Skewness measure as a complementary metric to address these shortcomings.

3.4.2 Scoring Emergent Scaling with Weighted Difference Skewness

The Weighted Difference Skewness $\gamma^*_{\Delta y}$ metric captures the symmetry of the distribution of metric differences. It is based on the Fisher-Pearson skewness coefficient, which is defined as the ratio between the third central moment of a distribution and the cube of the standard deviation. Letting $\Delta y_i = y_{i+1} - y_i$ denote consecutive performance deltas, the Weighted Difference Skewness, $\gamma^*_{\Delta y}$, is defined as:

$$\begin{aligned} \gamma^*_{\Delta y} &= \max(|\Delta y_i|) \cdot \gamma_{\Delta y} \\ \text{where } \gamma_{\Delta y} &= \frac{\mu_3(\Delta y)}{\sigma^3(\Delta y)} \end{aligned} \tag{4}$$

Here, $\mu_3(\Delta y)$ is the third moment of the metric difference distribution, $\sigma(\Delta y)^3$ is the cube of the standard deviation, and γ represents the adjusted Fisher-Pearson skewness coefficient[4]. **Importantly**, note the weighting factor $\max(|\Delta y_i|) = y_{max} - y_{min}$, which is the maximum observed metric difference—without this weighting factor, the skewness coefficient would be scale-invariant, and thus not sensitive to the magnitude of the metric change.

The intuition for using this metric is as follows: metric difference distributions concentrated at lower values but with a long right tail are characteristic of an emergent scaling curve—with mostly gradual increases and a small number of large jumps. This metric applies to datasets with both positive and negative Δy_i values, unlike the Breakthroughness$^+$ metric.

For both Breakthroughness$^+$ and Weighted Difference Skewness, a higher score indicates sharper, more abrupt scaling. It is important to bear in mind that these measures of emergent scaling are chosen for their mathematical properties and do not have a direct interpretation. Whilst being rooted in intuitive properties of emergent scaling—Breakthroughness$^+$ captures the ratio of the overall change in response variable y to the average change between consecutive points, Weighted Difference Skewness is the symmetry of the distribution of metric differences—they are not directly interpretable in isolation, and should primarily be used to *compare* the scaling behaviour between models, datasets, or individual samples.

4 Results

This section evaluates the scaling behaviour of the three reasoning models (Deepseek-R1-Distill-32B, QwQ-32B, Phi-4-Reasoning-Plus) across the three science and mathematics benchmarks (AIME24, AIME25, GPQA), with respect to the emergence scores introduced in Section 3—Breakthroughness$^+$ and Weighted Difference Skewness[5]. We begin with a comparison across datasets of trends in emergent scaling (Section 4.1), before focusing on AIME25, studying whether the same instances within the dataset consistently exhibit emergent scaling behaviour across models (Section 4.2)[6]. We then investigate the extent to which emergent test-time scaling differs across model families (Section 4.3), and finish by studying trends across model sizes, evaluating four models in the Deepseek-R1-Distill family (1.5B, 7B, 14B, 32B) across AIME25 (Section 4.4). Whilst this section focuses on the empirical findings of the study, the main implications are discussed further in Section 5.

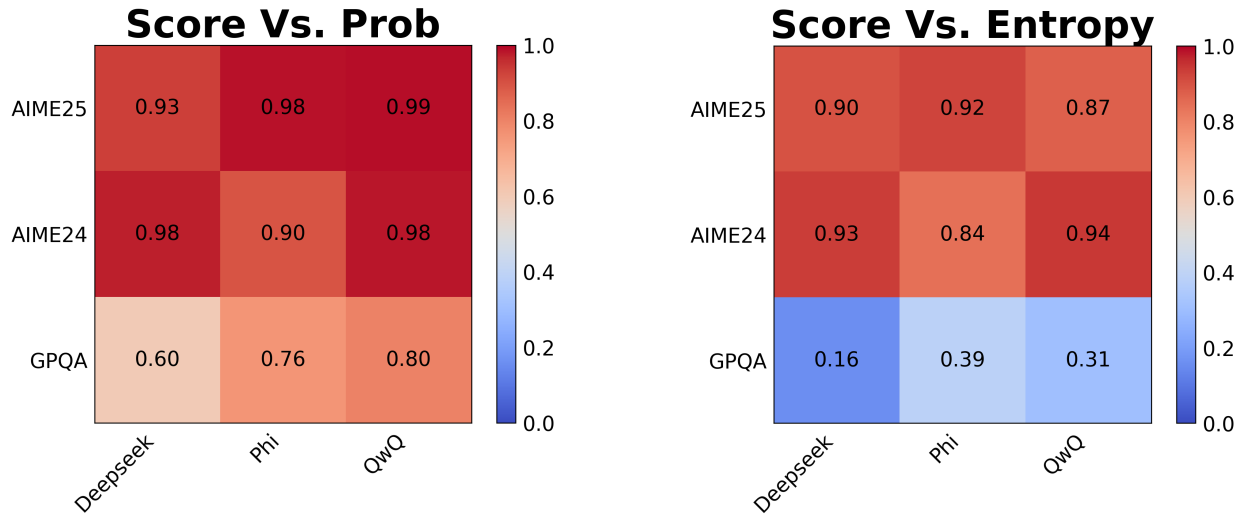
Before proceeding, we briefly highlight the correlation between the discrete and continuous metrics over the test-time compute budgets to highlight some key scaling trends. Recall that continuous metrics, in addition to discrete metrics, are evaluated to avoid the "mirage" effect of emergent scaling that results when only using discrete metrics (Schaeffer et al., 2023). The relationship between these metrics is shown in Figure 1, where the Pearson correlation coefficients between the accuracy (score) and ground truth probability, and accuracy and negative entropy (negentropy) are shown for all model-dataset pairings. Strong correlations are observed between accuracy and ground truth probability for all pairings; a similar result holds for the accuracy-negentropy correlations **except** for GPQA. The weak correlation between accuracy and negentropy for GPQA results from the saturation of the accuracy and probability metrics for larger test-time budgets, whilst negative entropy increases—in fact, for some instances in GPQA, a strong *negative* correlation is observed between these two metrics, implying that inverse scaling effects (Gema et al., 2025) are present. In other words, longer reasoning budgets for GPQA do not always improve the model's accuracy, but lead to increased

[4]The adjusted skewness coefficient includes the correction factor $\frac{\sqrt{N(N-1)}}{N-2}$ for bias correction in small samples.

[5]Throughout this section, the Breakthroughness+ and Weighted Difference Skewness scores are at times referred to as Breakthroughness and Skewness respectively

[6]Similar analyses for AIME24 and GPQA are left to Section A.1

model confidence, which if placed on the incorrect solution, will lead to a decrease in accuracy. AIME24 and AIME25 do not exhibit this inverse scaling effect.



(a) **(b)**

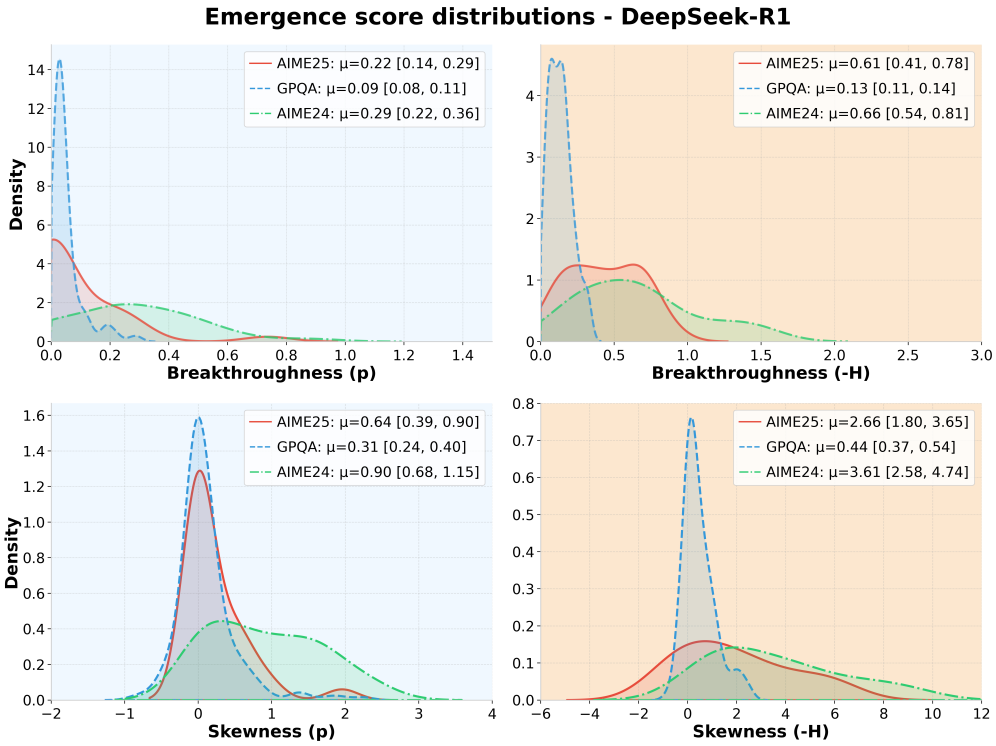
Figure 1: Pearson correlations between discrete (accuracy) and continuous (ground truth probability, negative entropy) metrics. As expected, strong correlations between the discrete and continuous metrics are observed for most model-dataset pairings. The weak correlation between accuracy and negentropy for GPQA indicates inverse test-time scaling effects.

4.1 Emergent test-time scaling varies substantially across datasets

We begin by investigating trends in emergent test-time scaling across different datasets. The research question (Section 1.1) addressed in this section is:
"To what extent does emergent scaling vary across different datasets? Are the returns to test-time compute best modelled by a log-linear relationship, or are there cases where other fits (such as piecewise linear) are more appropriate?"

This section studies the distribution of the emergence scores obtained by evaluating the scaling profiles of each instance across GPQA, AIME25, and AIME24 against the emergence scores Breakthroughness$^+$ and Weighted Difference Skewness. Here we analyse the scaling curves obtained with the Deepseek-R1-Distill-32B model (results for Phi-4-Reasoning-Plus and QwQ-32B are shown in Section A.1). Once again, the focus is on aggregated trends across datasets—an analysis of individual dataset instances displaying strong emergent scaling is deferred to Section 4.2. Note that high scores (large, positive values) when the probability metric is evaluated with Breakthroughness$^+$ or Skewness indicates that a model is more likely to see sharp, abrupt *progress towards the ground truth* answer, whilst high scores when the negentropy metric is evaluated with Breakthroughness$^+$ or Skewness indicates that a model is more likely to see sharp, abrupt increases in the *confidence* it assigns to a given solution (which may be or may not be the correct one).

Figure 2 shows the full distribution of emergence scores across the continuous metrics—ground truth probability and negative entropy, alongside 95% confidence intervals for the population means. Values for the emergence scores evaluated with respect to ground truth probability are coloured blue, and values for the scores when evaluated on the negentropy metric are coloured orange, to emphasise the conceptual difference between the two metrics. The sample means and 95% confidence intervals (for the population means) are reported for each dataset in the top right of the panel, with the confidence intervals calculated using bootstrap resampling with 1000 samples. The emergence score distributions for ground truth probability (blue) show that AIME24 displays the sharpest increases in model performance when evaluated with both Breakthroughness$^+$ and Weighted Difference Skewness, followed by AIME25, and then GPQA—this is evidenced by the green distribution shifted rightwards relative to the other two. GPQA displays the smoothest scaling behaviour, with Breakthroughness$^+$ scores in particular being clustered around 0. Similar trends hold for negative entropy (orange), indicating that model confidence increases at the sharpest rate when evaluated on AIME24.



	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
Skewness (p)	527	0.258	4092	2.3×10^{-5}	3516	0.016
Breakthroughness (-H)	501	0.455	5581	8.9×10^{-15}	4556	2.5×10^{-6}
Skewness (-H)	558	0.112	5083	1.5×10^{-13}	4051	4.1×10^{-5}

Figure 2: Full distributions of emergence scores when ground truth probability (p) and negative entropy ($-H$) scaling curves from Deepseek-R1-Distill-32B are with respect to Breakthroughness$^+$ and Weighted Difference Skewness. Sample means and 95% confidence intervals for population means (calculated using bootstrap resampling with 1000 samples) are displayed in the top right. Values for GPQA across all emergence scores are concentrated around 0, whereas AIME2024 and AIME2025 have modes shifted rightwards, indicating a greater degree of emergent test-time scaling in these two datasets. Distributions for ground truth probability (model performance) are shaded blue, with distributions for negentropy (model confidence) are shaded orange, to emphasise the conceptual difference between the metrics.

Table 1: Results of Mann-Whitney U tests for statistical significance between emergence score distributions between each pair of datasets, when evaluated with Deepseek-R1-Distill-32B. U is the test statistic, representing the number of times an instance from dataset 1 has a higher emergence score than an instance from dataset 2 (out of a total of $n_1 \times n_2$ instance pairs). At a significance level of $\alpha = 0.95$, there is a significant difference for all pairings of emergence score and metric between AIME24 and GPQA, 3/4 between AIME25-GPQA, and none between AIME24 and AIME25. This supports the finding that GPQA displays smoother scaling relative to AIME24 and AIME25, whilst the scaling behaviour between AIME24 and AIME25 is similar. Results for QwQ-32B and Phi-4-Reasoning-Plus-14B are given in Section A.1.

To test for statistical significance between the distributions across datasets, we conducted a Mann-Whitney U test for each pair of datasets. The results are shown in Table 1. AIME24 and AIME25 show no statistically significant difference across any pairing of emergence scores and metric, which is to be expected given the distributional similarity between the two datasets. However, AIME24 and GPQA see significant differences across all pairings of emergence

scores and metrics, and AIME25 and GPQA see significant differences across all pairings but one.[7] These findings suggest that AIME24 and AIME25 display much stronger emergent scaling behaviour relative to GPQA.

These findings can be validated by plotting the aggregate scaling curves over the full datasets—which is done in Figure 3. As expected, we see much sharper test-time scaling for the AIME datasets compared to GPQA at the aggregate level. The former two datasets see a noticeable increase in accuracy, ground truth probability, and negative entropy around the 2^{10} token budget, before reaching a final accuracy of 60% and 40% respectively at the 2^{13} token budget. The scaling behaviour is roughly piecewise linear—a key characteristic of emergent scaling behaviour. In contrast, GPQA sees a much smoother scaling curve, with accuracy, ground truth probability, and negative entropy increasing at a more gradual rate. Additionally, the accuracy and ground truth probability saturation for GPQA are clear for higher test-time budgets of 2^{12} and 2^{13} tokens, whilst the negative entropy continues to scale—this leads to the weaker correlations between accuracy and negentropy that was seen in Figure 1[8].

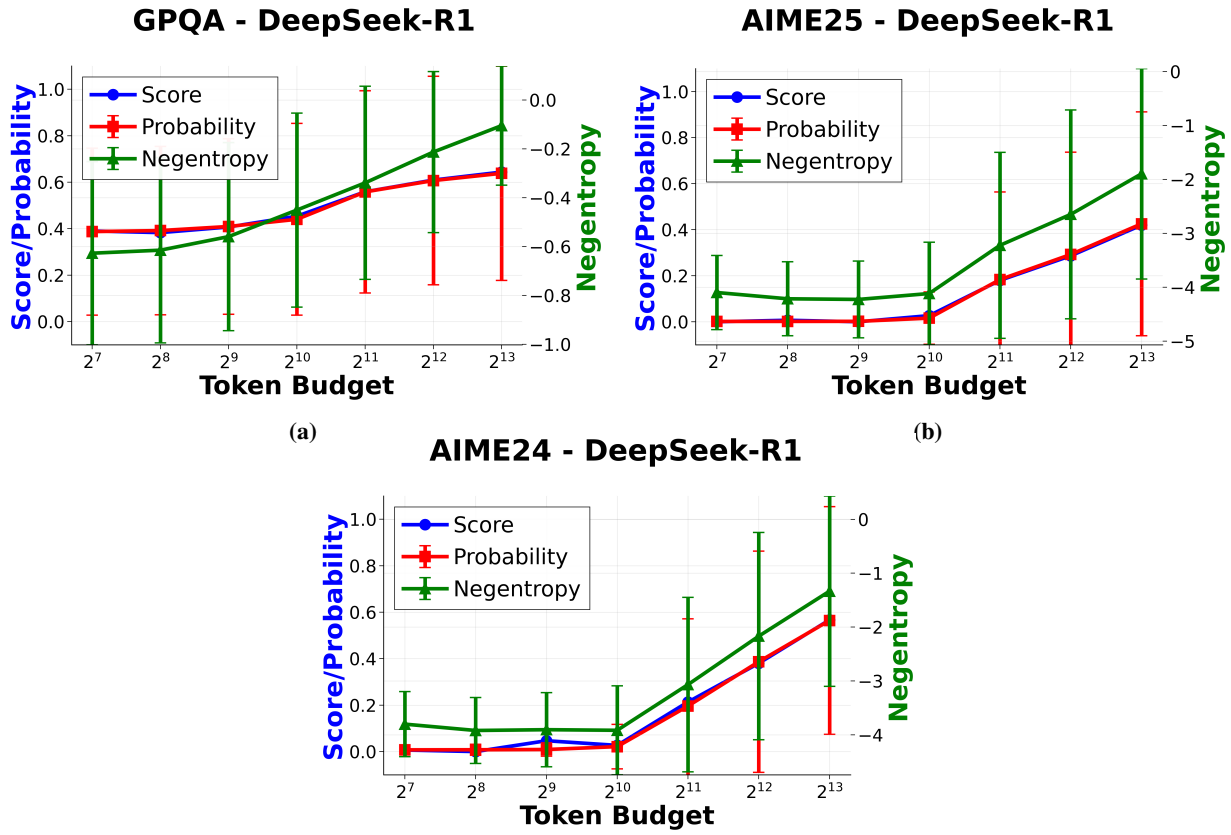


Figure 3: Aggregate scaling behaviour across all metrics (score, probability, negentropy) and datasets when evaluated with Deepseek-R1-Distill-32B. Even when aggregated across the full dataset, AIME2024 and AIME2025 display sharp, non-linear scaling (past 1024 tokens), whilst GPQA remains smoother. Error bars are found by taking the standard deviation across five runs. Note that negentropy scales differ across plots.

[7]Namely, the Breakthroughness$^+$ score evaluated on ground truth probability (p=0.137). However, we do not take this result as substantially weakening the broader trend; increased statistical power could resolve this discrepancy.

[8]The correlation between accuracy and negentropy for GPQA in Figure 3a does not appear to match the 0.16 Pearson correlation coefficient reported in Figure 1. This is due to how results are aggregated—in Figure 1, correlation coefficients are computed across the 198 instances in GPQA before being averaged (mean), whilst in Figure 3a, the scaling profiles are averaged across the 198 instances first. However, the main conclusion still holds—at higher token budgets, accuracy and negentropy disassociate due to inverse scaling effects.

4.2 Individual dataset instances consistently exhibit emergent test-time scaling

Whilst the previous section demonstrated that different datasets exhibit varying degrees of emergent scaling at the aggregate level, substantial variation can still exist *within* a given dataset. We look to further explore where emergent scaling results from dataset-level characteristics (such as problem domain and difficulty distribution) vs. instance-level characteristics (such as problem solution space size, number of reasoning steps required, and even prompt formatting). Therefore, this section investigates whether the same dataset instances (also referred to as problems) consistently exhibit emergent test-time scaling. The specific research question addressed in this section is:

"To what extent do individual dataset instances consistently exhibit emergent test-time scaling? Across multiple models, do the same instances consistently exhibit sharp, abrupt increases in performance?"

We focus our study on AIME25 as it showed substantial emergent test-time scaling in the previous section. Similar analyses for AIME24 and GPQA are presented in Section A.2. To begin, we evaluate the correlation across models between the ground truth probability emergence scores assigned to individual AIME25 problems. Specifically, each instance is first assigned a Breakthroughness$^+$ and Weighted Difference Skewness score for the ground truth probability scaling curve. Then, the instances are ranked by their emergence scores, with lower rankings corresponding to higher emergence scores. These rankings are aggregated amongst the Breakthroughness and Skewness scores[9], and the correlations of these rankings amongst the three model pairings (Deepseek-R1-Distill-32B vs Phi-4-Reasoning-Plus, Deepseek-R1-Distill-32B vs QwQ-32B, and Phi-4-Reasoning-Plus vs QwQ-32B) are found. This is equivalent to evaluating the Spearman correlations of the emergence scores across models. Results are shown in Figure 4. A strong correlation is seen across all model pairings, with a mean correlation coefficient of $\rho = 0.803$. Across all three models, the same instances consistently exhibit sharp, abrupt increases in ground truth probability; this is evidenced by the cluster of instances in the top-right of the figure. The presence of a strong correlation between emergence scores across models supports the hypothesis that features of the individual problem have a strong influence on the degree of emergent test-time scaling.

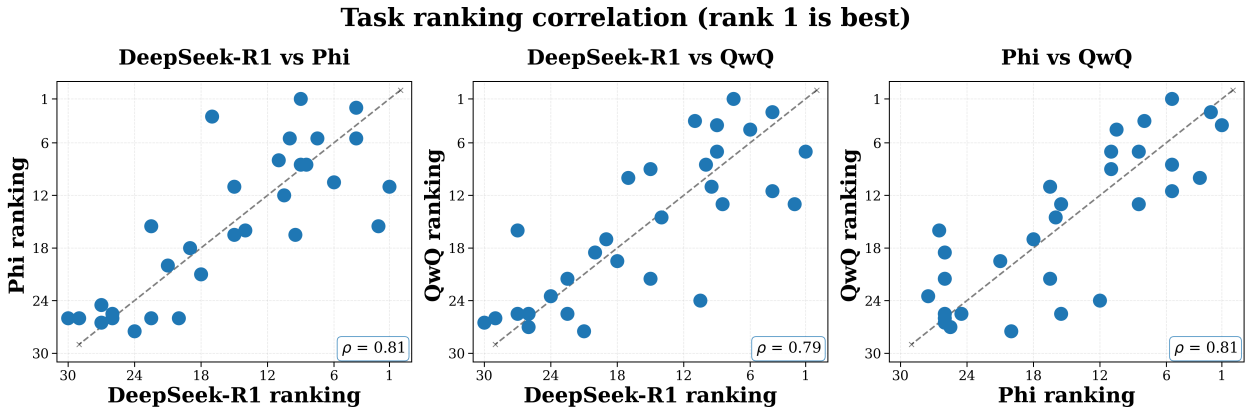


Figure 4: Spearman correlation of Skewness scores across AIME25 instances. For each model, instances are ranked by their emergence score—with rank 1 corresponding to the highest score—then the correlation coefficient of the rankings is calculated (shown in the bottom right of the panel). The strong correlation suggests that the input prompt has a significant effect on the likelihood of emergent test-time scaling. The same instances consistently see sharp, non-linear scaling across all models.

To better understand the scaling behaviour of these instances, we find the top-k instances displaying emergent test-time scaling when evaluated with Deepseek-R1-Distill-32B[10], and plot them in Figure 5. For AIME25, this corresponds to instances 2, 3, 15, and 5, which are given in Table 2. Observe that the scaling profiles across different models are

[9]This is done by simply adding the rankings. This has the effect of weighting the importance of Breakthroughness and Skewness scores equally.

[10]Specifically, we rank AIME25 instances with respect to the Breakthroughness$^+$ and Weighted Difference Skewness scores when evaluate with Deepseek (with lower rankings implying stronger emergent test-time scaling), aggregate these rankings, and choose the four (4) instances with the lowest rankings.

remarkably consistent; for example, instance 2 sees ground truth probability increase sharply from approximately 0.0 to 1.0 across a single doubling of thinking tokens when evaluated with Deepseek-R1-Distill-32B, Phi-4-Reasoning-Plus, and QwQ-32B. More generally, the number of doublings over which the ground truth probability increases from its initial value (0) to its final value (1) is consistent across the models, taking approximately 1-2 doublings. However, the exact token budget at which the abrupt scaling occurs varies across models. Observe that for instance 2, Deepseek and Phi see a notable improvement in ground truth probability (and negentropy) at 2^{10} tokens, but QwQ sees a jump in metrics at 2^{11} tokens. A similar pattern holds for instance 3 and 5. For dataset instance 15, Phi sees an increase in probability first at 2^9 tokens, then Deepseek at 2^{10} tokens, and finally QwQ at 2^{11} tokens. It is interesting to note that QwQ usually sees the most delayed onset of emergent test-time scaling, and this finding is discussed further in Section 5.1.

Similar results hold for AIME24 and GPQA. These are shown and discussed in Section A.2.

These findings support the hypothesis that instance-level features have a strong influence on the nature of test-time scaling observed, in contrast to dataset-level features. Note also that emergent scaling behaviour is observed across all three metrics (score, probability, and negentropy), providing strong evidence that the abrupt scaling is a feature of the model and tasks themselves, rather than being induced by metric choice. The observation that the onset of sharp increases in performance varies across models leads to the question of whether there are other differences in emergent test-time scaling between model families—we turn to this question in the next section.

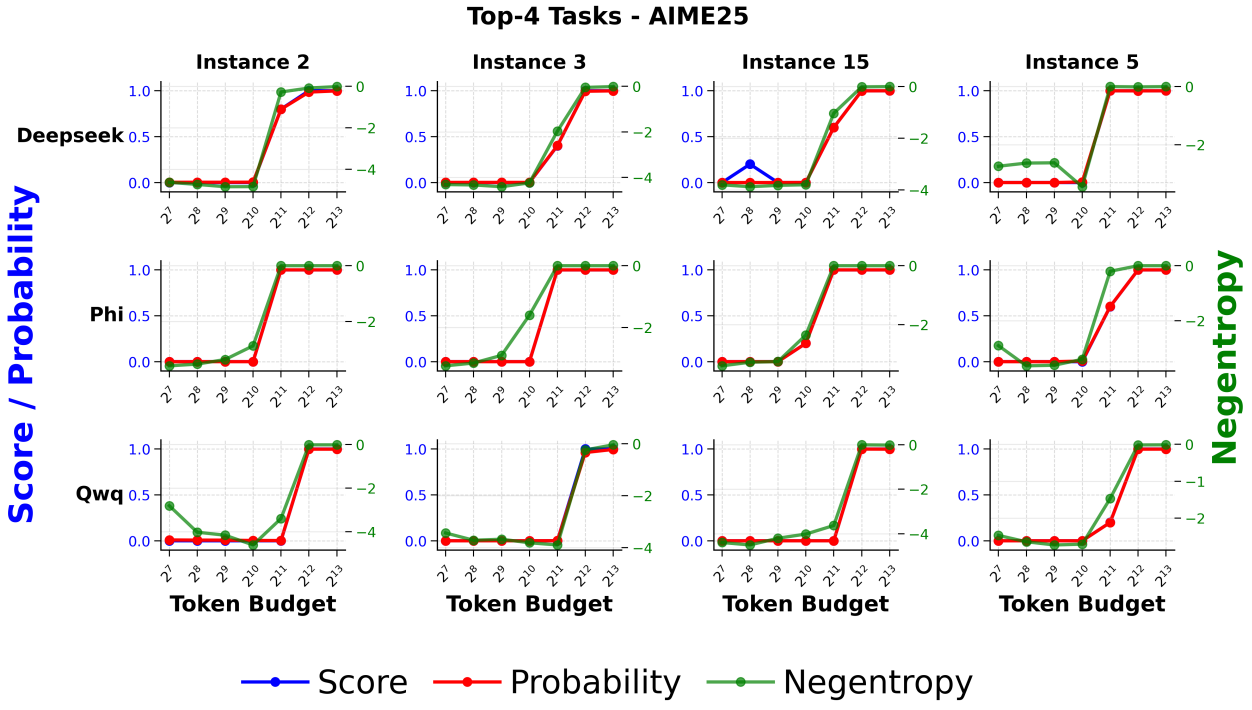


Figure 5: Top 4 samples in AIME25 dataset when ordered by emergence scores of Deepseek-R1-Distill-32B scaling curves. Scaling behaviour is consistent across models, with the same instances consistently exhibiting sharp, non-linear scaling behaviour. However, the test-time compute budget at which sharp scaling occurs can vary across models. For example, Instance 2 sees sharp scaling at approximately 2^{10} thinking tokens for Deepseek-R1 and Phi-4-Reasoning-Plus, but at approximately 2^{11} thinking tokens for QwQ.

Instance Index	Problem
2	The 9 members of a baseball team went to an ice-cream parlor after their game. Each player 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 was greater than the number of players who chose vanilla, which was greater than the number of players who chose strawberry. Let N be the number of different assignments of flavors to players that meet these conditions. Find the remainder when N is divided by 1000.
3	Find the number of ordered pairs (x, y), where both x and y are integers between -100 and 100 inclusive, such that $12x^2 - xy - 6y^2 = 0$.
15	Six points A, B, C, D, E, and F lie in a straight line in that order. Suppose that G is a point not on the line and that $AC = 26$, $BD = 22$, $CE = 31$, $DF = 33$, $AF = 73$, $CG = 40$, and $DG = 30$. Find the area of $\triangle BGE$.
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 $r^2 + s^2$.

Table 2: AIME25 top-k instances displaying emergent test-time scaling

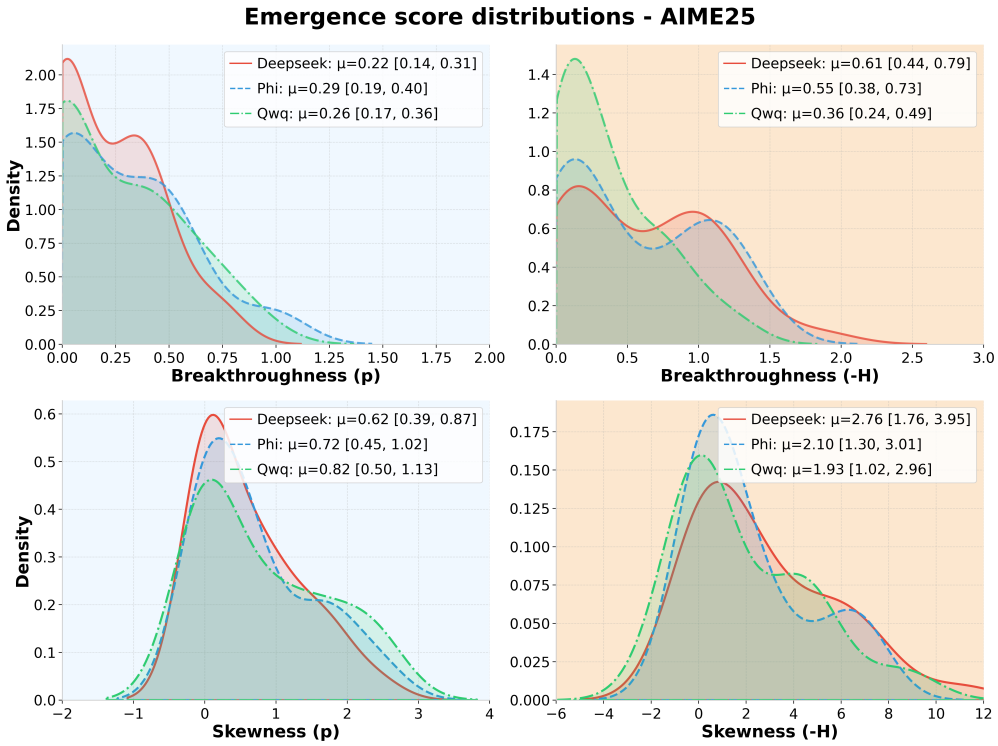
4.3 Emergent scaling is consistent across model families, but occurs at different token budgets

In this section, we investigate the question:
"To what extent does emergent scaling vary across different model families? Do some reasoning models display sharper scaling behaviour than others, or is this behaviour consistent across model families?"

Our study focuses on Deepseek-R1-Distill-32B, QwQ-32B, and Phi-4-Reasoning-Plus (14B), evaluating on the AIME25 dataset, with results for AIME24 and GPQA given in Section A.3. These three models achieve comparable performance on the AIME datasets[11]. Substantial differences in the degree of emergent test-time scaling between these models would indicate that the model training recipe (such as the method used for reasoning post-training (Kumar et al., 2025), model architecture, or training data mixture), influences the predictability of scaling behaviour, motivating further investigation.

Figure 6 shows the full distribution of emergence scores, alongside 95% confidence intervals for the population means across the three model families. Once again, ground truth probability distributions are shaded blue, and negentropy is shown in orange. Inspecting the distributions for ground truth probability, there is no clear difference across the three models. All three models have similar central values (mean) with overlapping confidence intervals, suggesting that the degree of emergent test-time scaling is consistent across model families. Regarding negative entropy distributions, QwQ shows a stronger concentration of Breakthroughness^{+} scores at low values, whilst all three models have similar distributions across negative entropy Skewness scores. Thus, a preliminary conclusion from observing these distributions is that there is no substantial difference in the degree of emergent test-time scaling across the Deepseek-R1-Distill, QwQ, and Phi-4-Reasoning-Plus model families.

[11] Deepseek-R1, QWQ, and Phi achieve pass@1 scores of 72.6%, 79.5%, and 81.3% respectively on **AIME24** (Guo et al., 2025; QwenTeam, 2025; Abdin et al., 2025). We were unable to find performance data for all three models on AIME25, but expect the two datasets to be similar in terms of difficulty.



	DeepSeek vs Phi $(n_1 = 30, n_2 = 30)$		**DeepSeek vs QwQ** $(n_1 = 30, n_2 = 30)$		**Phi vs QwQ** $(n_1 = 30, n_2 = 30)$	
Emergence Score	U	p	U	p	U	p
Breakthroughness (p)	413	0.590	436	0.842	461	0.877
Skewness (p)	469	0.785	462	0.865	441	0.900
Breakthroughness (-H)	469	0.785	579	0.058	546	0.158
Skewness (-H)	500	0.464	544	0.167	502	0.446

Figure 6: Full distributions of emergence scores across models when evaluated on AIME25. No model consistently exhibits the greatest degree of emergent test-time scaling across all emergence scores, evidenced by the relative similarity in distributions and corresponding population mean intervals.

Table 3: Results of Mann-Whitney U tests for statistical significance between emergence score distributions between each pair of model families, when evaluated on AIME25. U is the test statistic, representing the number of times an instance from model family 1 has a higher emergence score than an instance from model family 2 (out of a total of $n_1 \times n_2$ instance pairs). At a significance level of $\alpha = 0.95$, there is no statistically significant difference between any pairings of emergence score and metric between any model family pairings. This supports the finding that emergent scaling trends do not differ substantially between model families.

Mann-Whitney U tests are conducted across model pairs to check for statistical significance (or lack thereof) between the distributions. The results from this are shown in Table 3. As suggested from the visual inspection of the distributions in Figure 6, there are no statistically significant differences between the distributions across model pairs, implying that all model families exhibit similar degrees of sharp, abrupt scaling behaviour with respect to both ground truth probability and negentropy, when evaluated on the AIME25 dataset. (Similar results are observed for AIME24, and to a lesser extent, GPQA, in Section A.3.)

To verify these findings, we examine the aggregate scaling curves on AIME25 across the three models in Figure 7. The three models show similar scaling behaviour, with the ground truth probability showing a noticeable increase around $2^9 - 2^{10}$ tokens, and negentropy following a similar pattern. However, Phi-4 exhibits a noticeable increase in ground truth probability at a lower token budget than Deepseek or QWQ (2^9 tokens vs 2^{10} tokens), repeating a similar pattern that was observed in the instance-level analysis in Figure 5—namely, *the token budget at which emergent scaling occurs*

can vary across models, even if the global scaling patterns are consistent. This is not picked up in the analysis of the emergence score distributions in Figures 6 and Table 3, as the Breakthroughness^{+} and Weighted Difference Skewness measure the abruptness and magnitude of scaling behaviour but are invariant to the point at which this occurs.

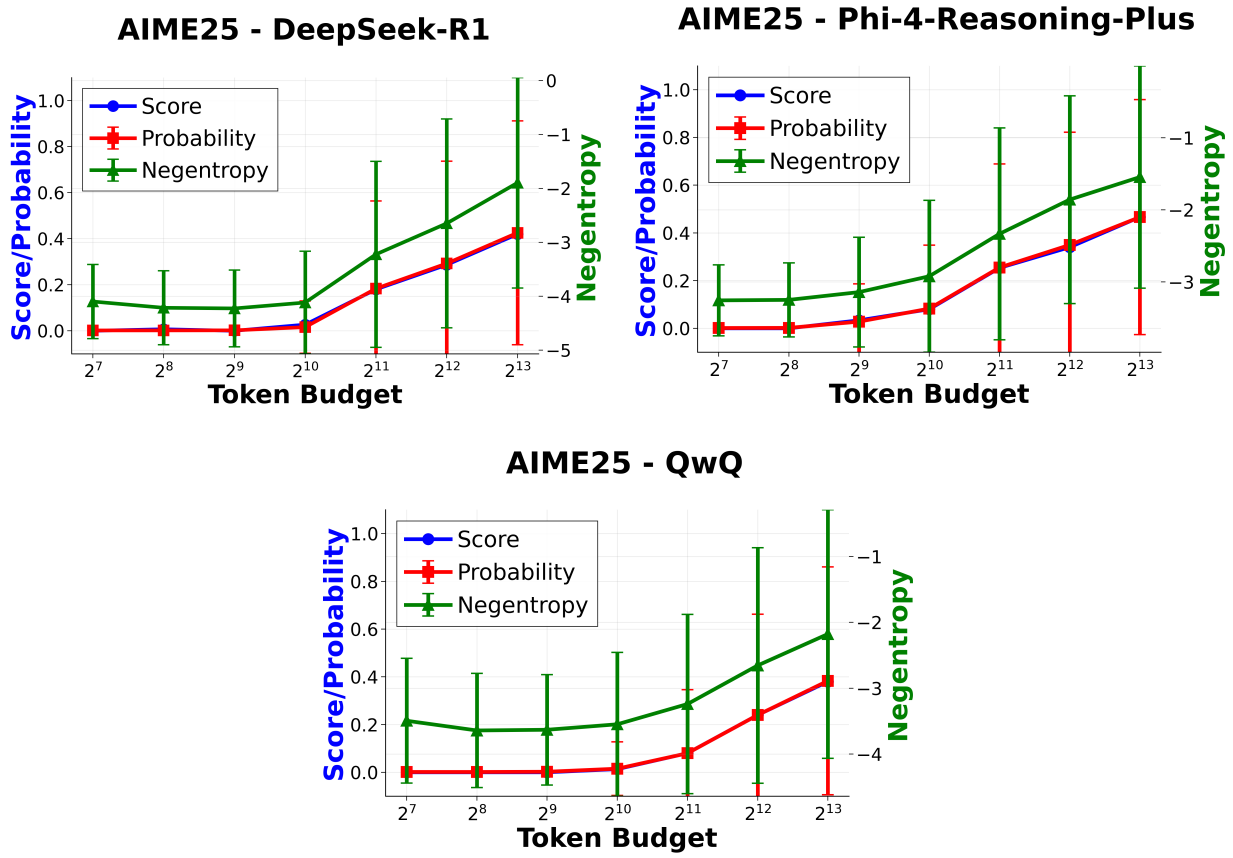


Figure 7: Aggregate scaling behaviour across Deepseek-R1-Distill, QwQ, and Phi-4-Reasoning-Plus models when evaluated on AIME25. Whilst the scaling profiles are mostly similar, featuring noticeably flat regions before an increase in accuracy/ground truth probability, Phi-4-Reasoning-Plus sees an increase in all metric values at lower token budgets (2^9 tokens, relative to 2^{10} tokens for the other two models).

Why might Phi-4-Reasoning-Plus see an earlier onset of emergent scaling? Several factors could explain this. The curation of high-quality, textbook-level data has been at the core of the Phi model family (Gunasekar et al., 2023), meaning that Phi-4-Reasoning-Plus may have been trained on higher-quality, less verbose reasoning traces. Additionally, the reinforcement learning phase to enhance the reasoning capabilities of Phi-4-Reasoning-Plus includes a specific length-aware accuracy score, which penalises excessively long reasoning traces (Abdin et al., 2025)—it is unclear if Deepseek-R1-Distill-32B or QwQ-32B have similar penalties in their objective functions. This trend should be investigated in future work, which is discussed in further detail in Section 5.1.

4.4 Larger models see sharper scaling of accuracy and ground truth probability

Finally, we investigate the relationship between model size and emergent test-time scaling. The research question addressed in this section is:

"To what extent does emergent scaling vary across model size (number of parameters)? Do larger models exhibit sharper scaling than smaller ones, or does model size have a negligible influence on emergent test-time scaling?"

We evaluate Deepseek-R1-Distill models at four sizes (1.5B, 7B, 14B, 32B) on the AIME25 dataset. Some key results for AIME24 are also discussed, with a more comprehensive overview of AIME24 results given in Section A.4. GPQA is omitted for this part of the study as it exhibits a lack of emergent test-time scaling at the dataset-level.

Figure 8 presents the full distributions of the emergence scores across model sizes for AIME25, with sample means and 95% confidence intervals for the population means in the top right of each panel. For ground truth probability, the density of Deepseek-R1-Distill-1.5B (red) dominates the distribution close to 0, suggesting that smaller models are less likely to see abrupt test-time scaling. On the other hand, the 14B and 32B models (green and purple, respectively) extend into the right of the distribution, with relatively little overlap in the confidence intervals for the population means when compared with the 1.5B model. Thus, larger models are more likely to exhibit sharp scaling of ground truth probability.

However, trends are less clear for the negative entropy plots—with the negentropy Skewness score distribution in particular showing relatively little difference between model sizes. Thus, the visual inspection of the distributions suggests that there is little variation in the degree of emergent scaling of negative entropy across model sizes.

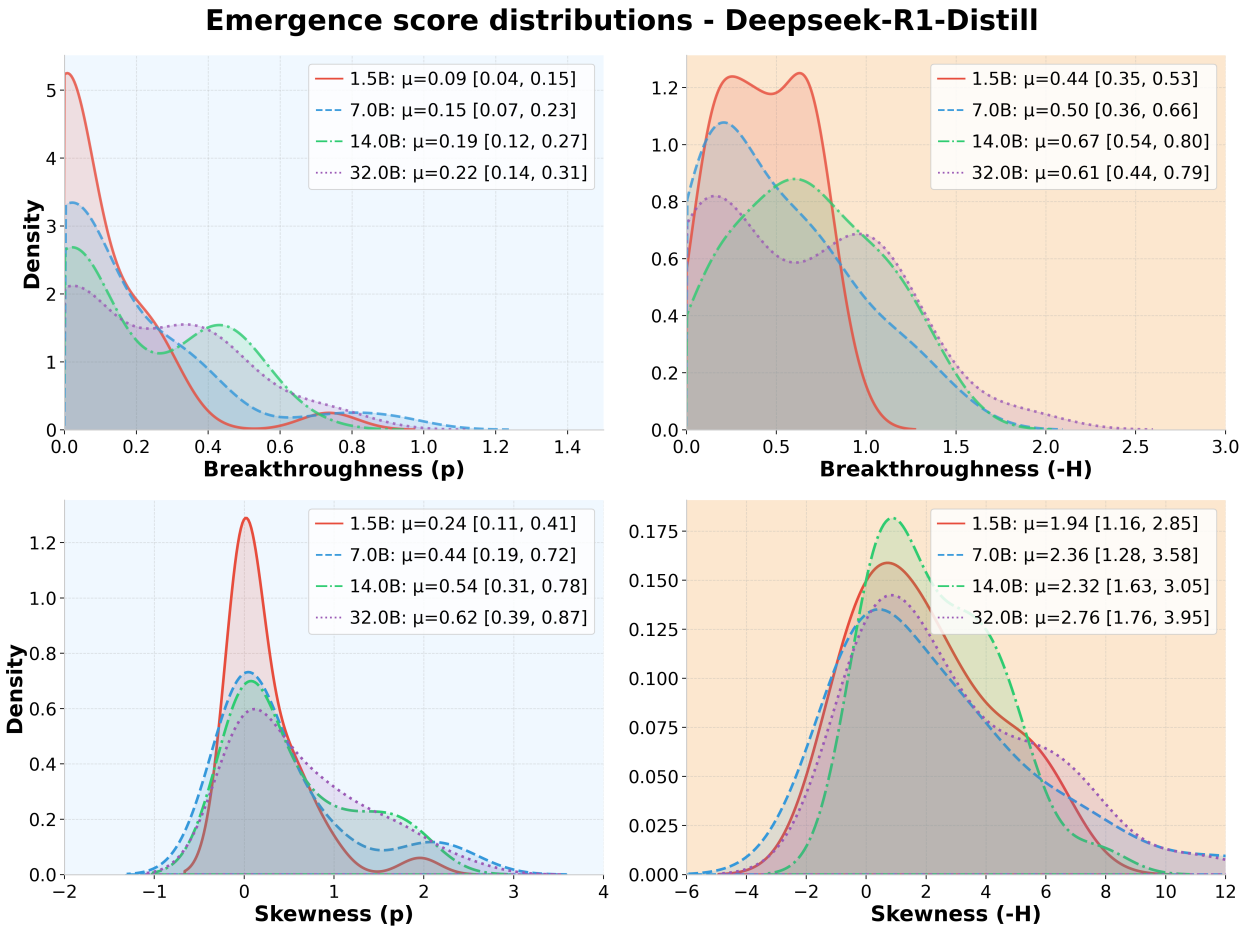


Figure 8: Full distributions of emergence scores across Deepseek-R1-Distill model size when evaluating on AIME25. Distributions for ground truth probability are shaded blue, distributions for negentropy are shaded orange. The 1.5B model dominates the distribution close to 0 for emergence scores on the ground truth probability metric, whilst the distributions of the 14B and 32B models extend to the right of the distribution. However, trends are less clear for emergence score distributions over negative entropy—there is little to distinguish the 7B, 14B, and 32B models in both the negative entropy distributions.

We explore this further in Figure 9, which presents summary statistics for the emergence score distributions for Deepseek-R1-Distill (1.5B, 7B, 14B, 32B), across AIME25. In particular, the sample means of the emergence scores are plotted, with error bars showing the standard error of the mean (SEM). The x-axis shows model size and is log-scaled.

The ground truth probability panels (left-hand side, blue) show the central values and error bar bounds increasingly monotonically for both Breakthroughness$^+$ and Skewness scores across all four model sizes. This adds further support to the idea that larger models are more likely to exhibit sharp, abrupt scaling behaviour in model performance, relative to smaller models. Additionally, the relationship is approximately linear, suggesting a consistent trend that could be extrapolated. The right-hand panels show a less consistent trend for emergent scaling of negentropy, echoing the visual inspection of the distributions in Figure 8. Here, central values and error bar bounds for the largest model (32B) are still greater than those for the smaller model (1.5B), but the trendline is not monotonically increasing. In fact, for negative entropy emergent scaling evaluated with Breakthroughness$^+$, the 14B model shows the sharpest scaling behaviour—though it is not clear if this is statistically significant.

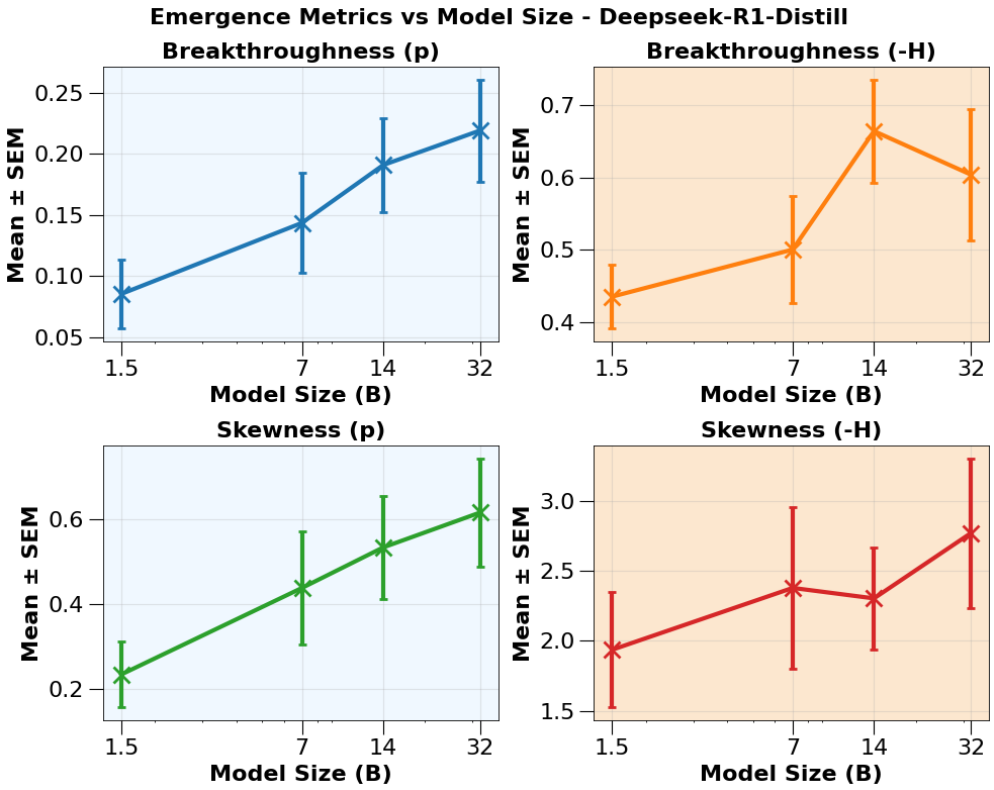


Figure 9: Summary statistics for the emergence scores across Deepseek-R1-Distill model size (1.5B, 7B, 14B, 32B) when evaluated on AIME25. **Blue panels**: When tracking the probability of ground truth metric, emergence scores (breakthroughness and skewness) monotonically increase with model size, indicating that larger models are more likely to see sharp, non-linear increases in ground truth probability. **Orange panels**: When tracking the negative entropy metric, emergence scores show a less consistent trend across model size; the global trend is still increasing but non-monotonic. This suggests that abrupt increases in model confidence across token budgets are not strongly correlated with model size.

However, it would be premature to conclude that emergent scaling of negative entropy does not display a consistent trend across model size. Inspecting the summary statistics of emergence distributions across model size for AIME24—shown in Figure 10—one can observe a much clearer increasing trend in the central values and error bar bounds for the negentropy metric evaluated with respect to Breakthroughness^{+} and Skewness. In this case, model confidence displays more abrupt increases with larger model sizes. The full score distributions are given in Section A.4, where there is a clearer distinction between the smaller and larger model distributions.

To briefly summarise, the results for emergent scaling trends across model size suggest that larger models are more likely to see sharp increases in ground truth probability (model performance); however, the case for emergent scaling in model confidence (measured by negative entropy) is less clear, with the AIME24 and AIME25 datasets displaying different trends.

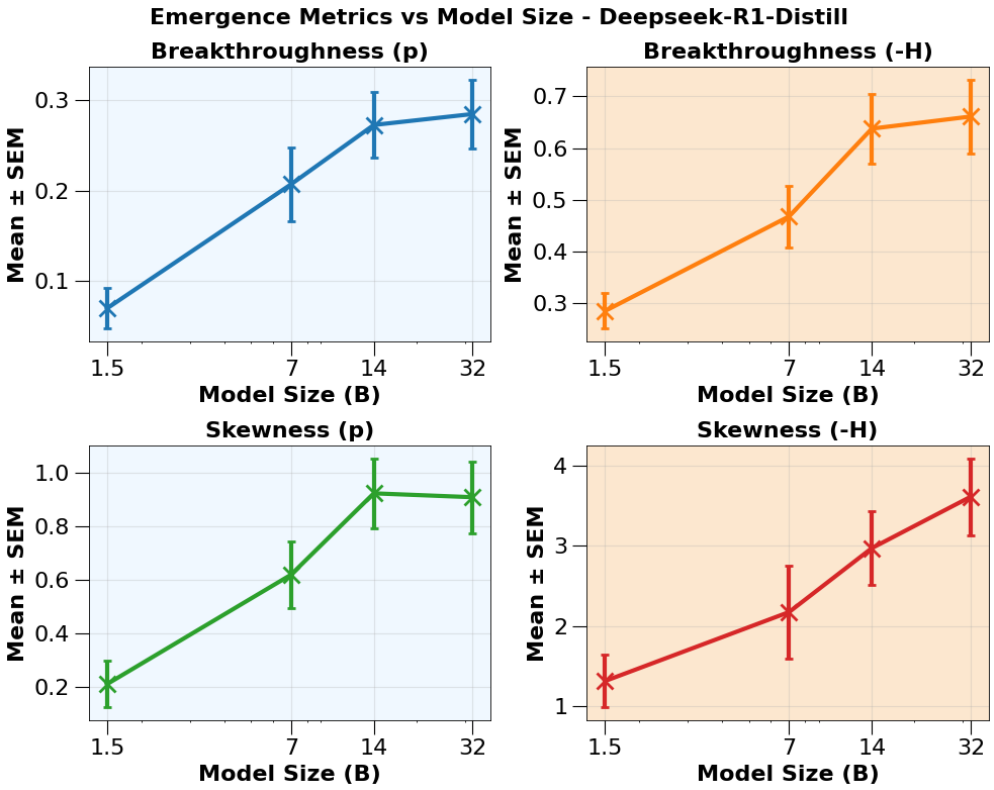


Figure 10: Summary statistics for the emergence scores across Deepseek-R1-Distill model size (1.5B, 7B, 14B, 32B) when evaluated on AIME24. **Blue panels**: When tracking the ground truth probability metric, emergence scores are generally increasing, but show signs of saturation at the 14B model size. **Orange panels**: When tracking the negative entropy metric, emergence scores increase across model size for AIME24. Additionally, these trends are more consistently increasing than for AIME25 (Figure 9). Further work is required to better understand test-time emergent scaling of model confidence with respect to model size.

We can verify the findings made for AIME25 by examining both aggregate and top-k scaling curves across the four model sizes for AIME25 (AIME24 results are given in Section A.4). Figure 11 shows scaling curves aggregated over all instances in AIME25. It is clear that the smaller models (1.5B and 7B) exhibit smoother scaling of performance metrics compared to the larger models—for example, the 1.5B model maintains 0 accuracy and ground truth probability until a reasoning budget of 2^{11} tokens, before it begins a gradual climb to a final accuracy of 20% at 2^{13} tokens. The 7B model sees an increase in accuracy at the 2^{10} budget mark, and climbs to a final accuracy of 30% at a reasoning budget of 2^{13}, showing some signs of saturation at the 2^{12} budget. In contrast, the 32B model sees a sharper increase in accuracy and ground truth probability at 2^{10} tokens, and reaches a final accuracy of 40% at 2^{13} tokens. The 14B model exhibits a similar pattern to the 32B model, but shows some signs of saturation at the 2^{12} budget, and reaches a slightly lower final accuracy. As discussed before, negative entropy profiles across model size have less consistent behaviour for AIME25. Interestingly, the 1.5B achieves higher negative entropy at 2^{13} reasoning tokens than the 32B model (-1.5 bits vs -1.9 bits), whilst having a much lower accuracy. (20% vs 40%). This shows the tendency for the smaller model to converge to *incorrect* solutions with equal, or higher confidence, than a larger model may have as it converges to the *correct* solution. This is an undesirable property—ideally, a model with lower accuracy would also display lower confidence. Further discussion of this point is given in Section 5.1.

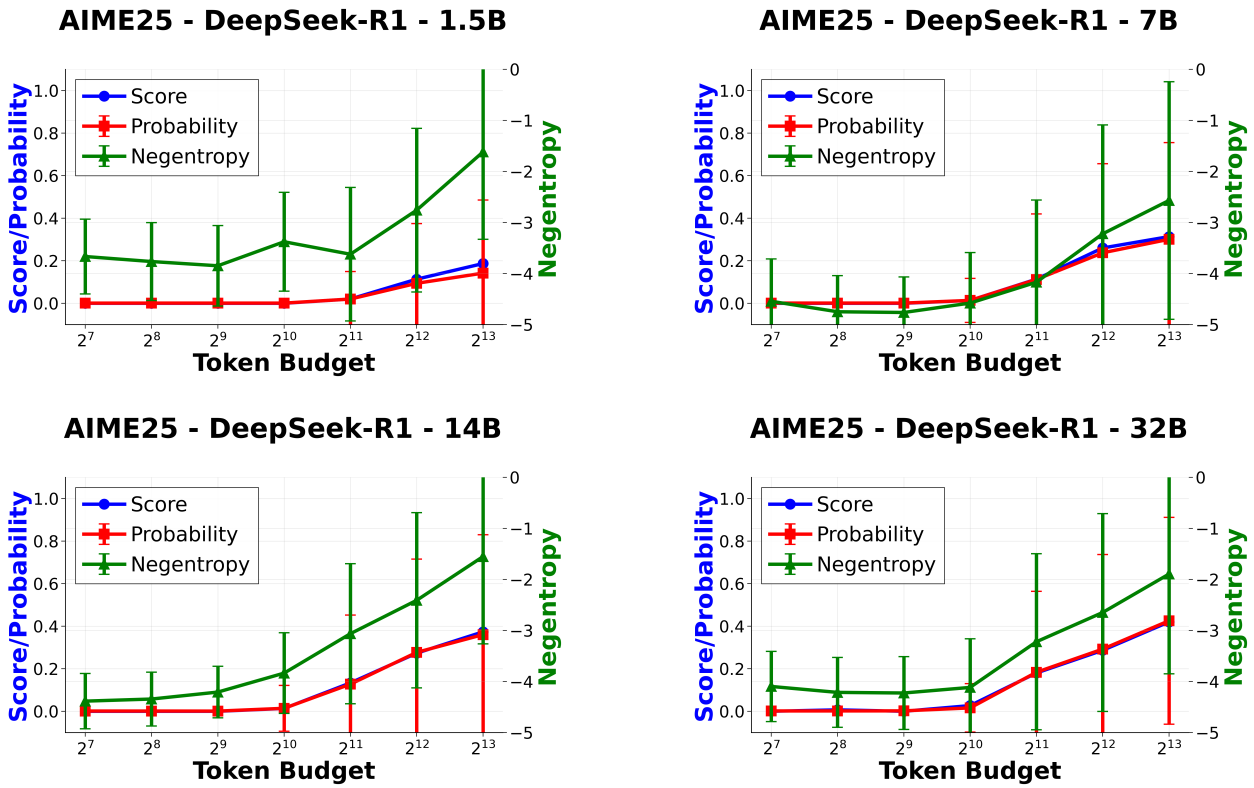


Figure 11: Aggregate scaling behaviour across Deepseek-R1-Distill model size (1.5B, 7B, 14B, 32B) when evaluated on AIME25. The 1.5B and 7B models show noticeably smoother scaling than the 32B model, and attain only 15% and 20% accuracy, respectively, relative to the 40% accuracy of the 32B model. Notice also the gradient of the score/probability metric curves at the point where the models start to make progress (2^{11} and 2^{10} tokens respectively); the 1.5B model shows a gradual increase in performance, whereas the 32B model shows a much sharper increase. This supports the finding that larger models see sharper scaling behaviour as test-time compute budget is scaled.

Finally, we plot the top-4 instances in AIME25 from Section 4.2 (instances 2, 3, 15, and 5—shown in Table 2) across model scales. The scaling profiles are shown in Figure 12. To begin, note the scaling patterns for instance 2. For the 1.5B model, accuracy and ground truth probability increase at 2^{12} tokens to 40% at 2^{13} tokens. For the 7B, 14B and 32B models, accuracy begins to increase at 2^{10} tokens, but the 14B and 32B models see sharper increases in ground truth probability at this budget than the 7B model. This matches the findings from the previous figures of larger models exhibiting sharper test-time scaling. Similar trends hold for the other three instances, with the smoother scaling of the 1.5B model being especially apparent. We also note that different model sizes see sharp scaling behaviour at different token budgets. In general, the larger the model, the earlier the increase in model accuracy begins—taking instance 3 for example, the 1.5B and 7B models begin seeing increases in accuracy and ground truth probability at 2^{11} tokens, whereas the 14B and 32B models see this begin at 2^{10} tokens.

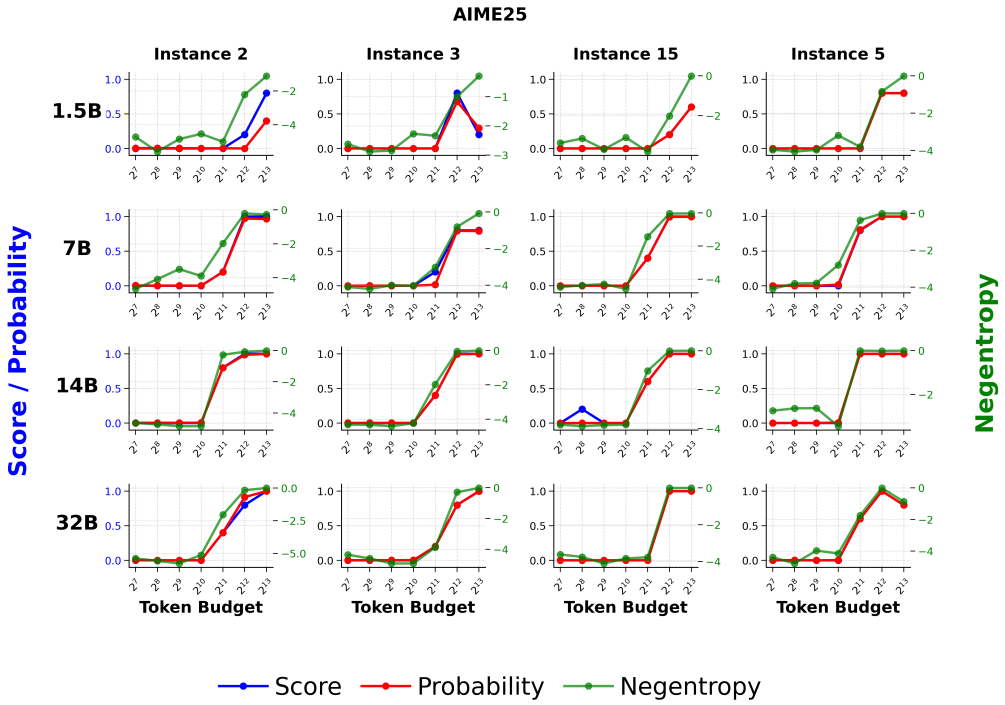


Figure 12: Scaling profiles of four samples identified as exhibiting emergent test-time scaling (see Section 4.2), across Deepseek-R1-Distill model size (1.5B, 7B, 14B, 32B). Model sizes are along rows, and samples are along columns. Sharper scaling profiles of ground truth probability are observed for larger models.

5 Discussion, Limitations, and Directions for Future Work

This section first highlights four key findings of our study, and discusses the implications of these in further detail (Section 5.1). This is followed by a discussion of the main limitations of this study (Section 5.2), and finally an outline of the key directions for future work (Section 5.3).

5.1 Implications of Key Findings

5.1.1 The relationship between performance and $\log(\textbf{compute})$ is not consistently linear

Previous work (Wu et al., 2025; Snell et al., 2024; Brown et al., 2025) found that model accuracy exhibits a linear relationship with respect to the logarithm of test-time compute budget. However, our results show that this is not always the case, and in certain cases, test-time scaling behaviour follows two-piece piecewise linear scaling instead. Caballero et al. (2023) fit piecewise linear curves to performance metrics as a function of training compute, model size, and training dataset size under the name of *broken neural scaling laws*; our analysis suggests that similar fits should be applied to test-time scaling behaviour when abrupt changes in performance are observed. The existence of critical token budgets (T_{crit}) at which abrupt increases in performance occur also has implications for the efficient allocation of compute at test-time. A growing body of work studies the *over-thinking* of reasoning models (Chiang and yi Lee, 2024; Chen et al., 2025; Sui et al., 2025), finding that models are prone to producing reasoning traces that are far longer than necessary to solve a simple problem. This study identifies several dataset instances that display rapid increases in accuracy and ground truth probabilities of ~0 to ~1 within a single doubling of test-time compute budget (See Section 4.2). Having a priori knowledge of the critical token budget, for example, i, T^i_{crit}, could be leveraged by allocating at most $2 \times T^i_{crit}$ tokens to the model, to benefit from this rapid transition whilst minimising overthinking. However, predicting the critical token budget for a given example appears non-trivial.

5.1.2 "Aha!" moments, breakthroughs in reasoning traces, and the influence of problem structure on emergent test-time scaling

The results of Section 4.2 identify key token budgets (referred to as threshold budgets, or T_{crit}) at which abrupt increases in performance are observed for a given dataset instance. This study focused on a quantitative analysis of this phenomenon, but did not carry out an investigation of qualitative patterns in dataset instances or reasoning traces that explain the sudden increase in performance—doing so would lead to a better understanding of the mechanisms underlying emergent test-time scaling. Many hypotheses could explain such behaviour, including:

1. The problem in question being highly compositional—if a problem is cleanly decomposable into multiple subproblems, a model might use up to T_{crit} tokens to reason about these subproblems, before beginning to piece these together to make progress towards the solution of the main problem after the T_{crit} threshold. This would lead to a scaling curve that is characteristic of emergent test-time scaling. This hypothesis aligns with the results of Shojaee* et al. (2025), which find that reasoning model accuracy collapses as problem complexity—measured by the number of sequential operations needed to solve a problem—increases.
2. The model makes a sudden breakthrough as it reasons, similar to moments of insight ("Aha!" moments) exhibited by humans (Kounios and Beeman, 2009), and purportedly, in previous studies of reasoning models (Guo et al., 2025). In human studies, Kounios and Beeman (2009) describe these moments as resulting in the sudden reinterpretation of a situation, which facilitates a new approach to the problem. Whilst care must be taken not to anthropomorphise the model's computations, our results warrant the further investigation of model reasoning processes around the T_{crit} budget—say, by investigating whether there are discontinuous changes in the model's latent representation of the input at this point.

The compositional problem hypothesis may be particularly worth exploring due to the strong dependence that emergent test-time scaling exhibits on individual dataset instances. Instances 2, 3, 15, and 5 of AIME25 consistently displayed sharp, abrupt increases in performance (see Section 4.2) independent of the reasoning model used—an analysis of

whether these instances had similar features or covered similar mathematical concepts (such as algebra, geometry, number theory, etc.) would provide insight into this.

5.1.3 Emergent scaling of ground truth probability increases across model size; Negative entropy scaling remains unclear

An interesting finding of Section 4.4 is that ground truth probability shows sharper scaling as model size is increased, but negative entropy results were inconclusive when studying AIME24 and AIME25. In particular, the data from AIME24 suggested that larger models display more abrupt increases in confidence than smaller models, whereas AIME25 results indicated that there was no substantial difference in the emergent scaling trends of negative entropy across model size. The former result is more desirable from a calibration perspective; ideally, models that often return incorrect responses should reflect this with higher uncertainty over possible solutions. However, this was not the case for AIME25—Figure 11 shows that the 1.5B model had similar final levels of negative entropy to the 32B model, despite the former having noticeably lower accuracy.

Why might it be the case that final accuracy and model confidence are more weakly correlated on AIME25 than they are on AIME24? One possible explanation is data leakage—Deepseek-R1-Distill-32B, QwQ-32B, and Phi-4-Reasoning-Plus were all released roughly one year after the AIME24 competition, increasing the likelihood of the models having seen these problems in the training data. If this is the case, then the test-time scaling behaviour of negative entropy across model size on AIME25 should be given more weight. Another explanation is that the nature of problems in AIME25 may be more prone to lead models to report incorrect solutions with high confidence, whereas AIME24 problems might offer more opportunities for internal verification—through sanity checks or by producing verifiable intermediate results. More opportunities for verification mean that models that follow an incorrect path are better able to identify this and reduce confidence in incorrect solutions. In contrast, problems that do not have checkpoints provide no feedback when exploring an incorrect path, leading models to confidently report incorrect answers. However, AIME24 and AIME25 are drawn from the same competition, only one year apart, and will thus have substantial distributional overlap. Additionally, this empirical result may simply be a finite sample size effect.

5.1.4 Variation across model families in the onset of emergent scaling

Section 4.3 found little change in the abruptness or magnitude of emergent scaling across model families, but differences in the token budget at which the sharp scaling occurred across the $2^9 - 2^{11}$ token range, with Phi-4-Reasoning-Plus (14B) exhibiting the earliest onset, followed by Deepseek-R1-Distill (32B), and then QwQ-32B. The reasons for the delayed progress displayed by QwQ-32B were not investigated in this study, but this is suggestive of variations in the approaches taken by each model to solve the problem. Perhaps QwQ-32B is more verbose in its traces, working through each individual arithmetic step, whilst Phi-4-Reasoning-Plus is more concise and direct, making more implicit assumptions in its reasoning. Another exploration is that some models may be more prone to "exploring" before "exploiting"—that is, QwQ models may be more willing to search over a large number of possible methods and solutions before committing to a chosen path, whereas Phi-4-Reasoning-Plus quickly makes a decision on how to proceed. An interesting avenue for future work would be to systematically study reasoning styles between different model families, to see if this can explain the differences in the onset of emergent scaling behaviour identified in this study.

5.2 Limitations

This study has the following limitations:

Only sequential scaling is used to increase test-time compute budget: There are many ways to utilise compute at inference-time, some of which are covered in Section 2.1.1. This study scaled test-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 and 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) shows signs of saturation on GPQA at the test-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 test-time compute (after having passed the critical token budget 2^{10} tokens). One key scaling trend that would set in at higher budgets are inverse scaling effects (Gema et al., 2025), where models exhibit negative returns to additional test-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 test-time scaling occurs for a broader range of tasks than those evaluated in this study.

5.3 Directions for Future Work

The implications of this work discussed in Section 5.1, as well as limitations identified in Section 5.2, suggest the following avenues for future work:

- Perform systematic analyses of reasoning traces around the T_{crit} critical budgets to investigate the mechanisms that underlie emergent test-time scaling. Such work could use existing structural breakdowns of reasoning traces (Marjanović et al., 2025) and investigate whether there are statistically significant differences in the structure of reasoning traces that lead to smooth versus abrupt scaling. Manual examination of reasoning traces could also be done, but would be time-consuming due to the length of the traces. A less interpretable approach would be to create a low-dimensional embedding of the reasoning trace and train a regression model to predict the Breakthroughness^{+} metric and Weighted Difference Skewness metric; if an accurate model can be trained, this would be suggestive of key reasoning trace features that determine the degree of emergent test-time scaling.

- Perform qualitative analyses of dataset instances that exhibit strong emergent test-time scaling. This could be done by collecting more instances that exhibit emergent scaling (i.e: across a broader number of datasets, and test-time scaling methods) and doing a large-scale analysis of instance features. One hypothesis to test against is the compositional problem hypothesis (Arora and Goyal, 2023; Barak, 2023), where tasks that are composed of multiple subtasks are more likely to exhibit abrupt scaling behaviour.

- Extend the study to parallel scaling methods, to determine whether emergent test-time scaling is observed as the number of sampled solutions is increased, or whether hyperparameters of the parallel scaling method (such as branching factor, reward model type, etc.) exhibit thresholding behaviour.

- Extend the study to a broader range of reasoning-heavy benchmarks in domains outside mathematics and science, such as coding (e.g: LiveCodeBench (Jain et al., 2024)) and strategic game playing (e.g: GameBench (Costarelli et al., 2024)), to see if some tasks are more likely to see emergent scaling than others.
- Increase the length of reasoning traces past 2^{13} tokens to investigate inverse scaling effects (Gema et al., 2025), most notably whether sharp, abrupt *decreases* in performance can be observed as test-time budget is scaled.

6 Conclusion

This study investigated the extent to which sharp, abrupt increases in language model performance metrics occur as test-time compute is scaled (emergent test-time scaling). Existing work has previously suggested a log-linear relationship between test-time compute and model accuracy; however, this was conducted for a relatively narrow range of benchmarks, and at a coarse level across several tasks, which can easily mask sharp scaling behaviour. Understanding the scaling behaviour between test-time compute and model performance is needed to improve scientific understanding of language reasoning models, as well as to optimally allocate test-time compute per example. Additionally, the presence of abrupt increases in model performance as test-time compute is scaled is suggestive of breakthrough moments in reasoning, akin to "Aha!" moments in human cognition (Kounios and Beeman, 2009), indicating that the model undergoes discontinuous changes in its interpretation of a problem whilst thinking. This study aimed to shed light on the frequency and influencing factors of such scaling behaviour.

Emergent test-time scaling was achieved through sequential scaling—controlling the length of reasoning traces by appending special continuation and force-stopping tokens (Muennighoff et al., 2025). This was done across three mathematics and science benchmarks—AIME24, AIME25, and GPQA-Diamond, and three model families—Deepseek-R1-Distill, QwQ, and Phi-4-Reasoning-Plus. Across test-time compute budgets, accuracy, ground truth probability, and negative entropy were tracked; these comprise a mixture of discrete and continuous metrics to assess model performance and confidence in solutions. The degree of emergent test-time scaling in these scaling curves was identified with two complementary metrics: Breakthroughness^{+}, which captures the ratio of the total change in a metric to the average change, and Weighted Difference Skewenss, which quantifies the symmetry of the distribution of metric differences. Within this experimental setup, the following questions were investigated, with the main findings given in red:

Q1: To what extent does emergent scaling vary across different datasets? Are the returns to test-time compute best modelled by a log-linear relationship, or are there cases where other fits (such as piecewise linear) are more appropriate?

A: There is substantial variation in emergent test-time scaling across datasets. In this study, AIME24 and AIME25 showed abrupt increases in performance around token budgets of $2^9 - 2^{10}$ tokens, whilst GPQA displayed more gradual scaling across all token budgets. This suggests that more flexible alternatives to log-linear fits are needed to model test-time scaling behaviour.

Q2: To what extent do individual dataset instances consistently exhibit emergent test-time scaling? Across multiple models, do the same instances consistently exhibit sharp, abrupt increases in performance?

A: The same dataset instances consistently exhibited emergent test-time scaling, across all reasoning models used for evaluation. For example, instances 2, 3, 15, and 5 from the AIME25 dataset all displayed sharp increases in the probability metric, from approximately 0.0 to 1.0, over 1-2 doublings of token budget. This suggests that features of individual problems have a strong influence on the degree of emergent scaling, in contrast to dataset-level characteristics.

Q3: To what extent does emergent scaling vary across different model families? Do some reasoning models display sharper scaling behaviour than others, or is this behaviour consistent across model families?

A: The abruptness and magnitude of emergent scaling across model families is consistent, with models showing little variation over the absolute change in metric values, and the number of compute budget doublings over which this occurs. **However**, the token budget at which abrupt scaling occurs varies across models. In our study,

Phi-4-Reasoning-Plus generally sees the earliest onset of emergent scaling, followed by Deepseek-R1-Distill, and then QwQ-32B.

Q4: To what extent does emergent scaling vary across model size (number of parameters)? Do larger models exhibit sharper scaling than smaller ones, or does model size have a negligible influence on emergent scaling?

A: Larger models show sharper scaling of ground truth probability than smaller models; in our study, Deepseek-R1-Distill 14B and 32B saw more abrupt increases in probability than the 1.5B and 7B models over the same number of token budget doublings. However, our results were inconclusive for emergent scaling trends of the negative entropy metric across model sizes, with AIME25 showing a weak correlation between model size and degree of negative entropy emergent scaling, whereas AIME24 showed a more consistent increasing trend.

Q5: Is the presence of emergent test-time scaling contingent on the metric used for evaluation? Does emergent scaling only occur when using discrete evaluation metrics, or are they present when tracking continuous evaluation metrics too?

A: Abrupt increases in all three metrics—accuracy, ground truth probability, and negative entropy—were consistently observed in this study, and were often correlated. This suggests that the observed scaling behaviours are features of the reasoning models and the tasks themselves, rather than being an artefact of metric choice.

These results open avenues for future work to investigate the mechanisms that underlie emergent test-time scaling. Alongside addressing the limitations of this study (Section 5.2), future studies should investigate the features of individual dataset instances that display strong emergent scaling—with the compositional complexity of the problem being a natural place to begin. Additionally, the influence of reasoning traces on emergent scaling should also be studied, to see if key "thought" patterns in language reasoning models lead to abrupt increases in performance, or whether there are discontinuous changes in the model's latent representation of the input at the critical token budget at which sharp scaling occurs.

Overall, this study contributes to the growing body of work on understanding the relationship between using additional test-time compute and model performance. The positive results for emergent scaling provide evidence in the direction that language reasoning models may also exhibit "Aha!" moments in reasoning, akin to moments of breakthrough insight in humans. Should this be true, it would suggest that scaling test-time compute further could unlock qualitatively different abilities that have yet to be observed with current models.

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. (2025). Phi-4-reasoning technical report.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alabdulmohsin, I., Neyshabur, B., and Zhai, X. (2022). Revisiting neural scaling laws in language and vision.

Anderson, P. W. (1972). More is different: broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Arora, S. and Goyal, A. (2023). A theory for emergence of complex skills in language models.

Barak, B. (2023). Emergent abilities and grokking: Fundamental, mirage, or both? Blog post on Windows on Theory.

Berti, L., Giorgi, F., and Kasneci, G. (2025). Emergent abilities in large language models: A survey.

Brown, B., Juravsky, J., Ehrlich, R. S., Clark, R., Le, Q. V., Re, C., and Mirhoseini, A. (2025). Large language monkeys: Scaling inference compute with repeated sampling.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Caballero, E., Gupta, K., Rish, I., and Krueger, D. (2023). Broken neural scaling laws.

Chen, X., Xu, J., Liang, T., He, Z., Pang, J., Yu, D., Song, L., Liu, Q., Zhou, M., Zhang, Z., Wang, R., Tu, Z., Mi, H., and Yu, D. (2025). Do not think that much for 2+3=? on the overthinking of o1-like llms.

Chiang, C.-H. and yi Lee, H. (2024). Over-reasoning and redundant calculation of large language models.

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

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

Du, Z., Zeng, A., Dong, Y., and Tang, J. (2025). Understanding emergent abilities of language models from the loss perspective.

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

Feng, X., Wan, Z., Wen, M., McAleer, S. M., Wen, Y., Zhang, W., and Wang, J. (2024). Alphazero-like tree-search can guide large language model decoding and training.

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

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. (2025). Inverse scaling in test-time compute. *arXiv preprint arXiv:2507.14417*.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., and Li, Y. (2023). Textbooks are all you need.

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

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., et al. (2024). 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)*, pages 3828–3850.

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

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset.

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. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Hu, S., Liu, X., Han, X., Zhang, X., He, C., Zhao, W., Lin, Y., Ding, N., Ou, Z., Zeng, G., et al. (2023). Predicting emergent abilities with infinite resolution evaluation. *arXiv preprint arXiv:2310.03262*.

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

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

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

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

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

Kumar, K., Ashraf, T., Thawakar, O., Anwer, R. M., Cholakkal, H., Shah, M., Yang, M.-H., Torr, P. H. S., Khan, F. S., and Khan, S. (2025). Llm post-training: A deep dive into reasoning large language models.

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

Lubana, E. S., Kawaguchi, K., Dick, R. P., and Tanaka, H. (2024). A percolation model of emergence: Analyzing transformers trained on a formal language.

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

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

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

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

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. (2021). Show your work: Scratchpads for intermediate computation with language models.

Okawa, M., Lubana, E. S., Dick, R. P., and Tanaka, H. (2024). Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task.

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. (2024). Openai o1 system card.

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

QwenTeam (2025). Qwq-32b: Embracing the power of reinforcement learning.

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

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

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

Shojaee*, P., Mirzadeh*, I., Alizadeh, K., Horton, M., Bengio, S., and Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. In *NeurIPS*.

Snell, C., Lee, J., Xu, K., and Kumar, A. (2024). Scaling llm test-time compute optimally can be more effective than scaling model parameters.

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. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

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

Sui, Y., Chuang, Y.-N., Wang, G., Zhang, J., Zhang, T., Yuan, J., Liu, H., Wen, A., Zhong, S., Zou, N., Chen, H., and Hu, X. (2025). Stop overthinking: A survey on efficient reasoning for large language models.

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

Villalobos, P. (2023). Scaling laws literature review. *Published online at epochai. org*.

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

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

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

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

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022b). Chain of thought prompting elicits reasoning in large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

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

Wu, Y., Sun, Z., Li, S., Welleck, S., and Yang, Y. (2025). Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models.

Xu, F., Hao, Q., Zong, Z., Wang, J., Zhang, Y., Wang, J., Lan, X., Gong, J., Ouyang, T., Meng, F., et al. (2025). Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

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

Zhang, Q., Lyu, F., Sun, Z., Wang, L., Zhang, W., Hua, W., Wu, H., Guo, Z., Wang, Y., Muennighoff, N., King, I., Liu, X., and Ma, C. (2025). A survey on test-time scaling in large language models: What, how, where, and how well?

Zhang, Y. and Math-AI, T. (2024). American invitational mathematics examination (aime) 2024.

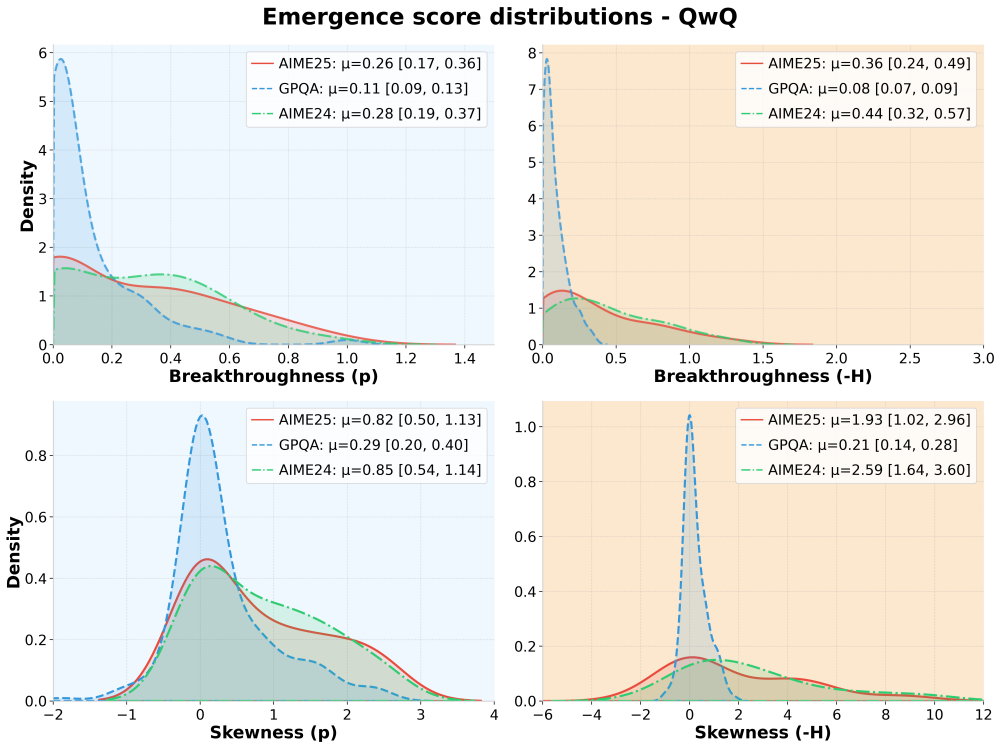
Zhang, Y. and Math-AI, T. (2025). American invitational mathematics examination (aime) 2025.

A Additional experimental results

This appendix provides additional experimental results that were not included in the main text of the thesis.

A.1 Additional emergence score distributions across datasets

This section contains additional results for emergence score distributions across datasets (AIME24, AIME25, GPQA) for the models QwQ-32B and Phi-4-Reasoning-Plus. The main text discussed results for Deepseek-R1-Distill-32B (Section 4.1). The main takeaway of Section 4.1—that, at the aggregate level, AIME24 and AIME25 exhibit strong degrees of emergent scaling whilst GPQA shows much smoother returns to test-time compute—also holds for QwQ-32B and Phi-4-Reasoning-Plus.



	AIME 2024 vs AIME25 ($n_1 = 30, n_2 = 30$)		AIME 2024 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)	489.00	0.5692	3915.00	0.005007	3499.00	0.1163
Skewness (p)	480.00	0.6627	3659.00	0.0007386	3427.00	0.009158
Breakthroughness (-H)	536.00	0.2062	5203.00	3.26×10^{-11}	4588.00	1.53×10^{-6}
Skewness (-H)	529.00	0.2458	4223.00	1.03×10^{-7}	3396.00	0.01027

Figure 13: Emergence score distribution across datasets for **QwQ-32B**. Sample means are shown in the top-right of each panel, alongside 95% confidence intervals for the population means (calculated using bootstrap resampling with 1000 samples).

Table 4: Results of Mann-Whitney U tests for the emergence score distributions across datasets for **QwQ-32B**. Similar to the results for Deepseek-R1-Distill-32B (Section 4.1), nearly all differences between AIME datasets and GPQA are statistically significant, whilst there is no significant difference between AIME24 and AIME25 (at significance level $\alpha = 0.95$).

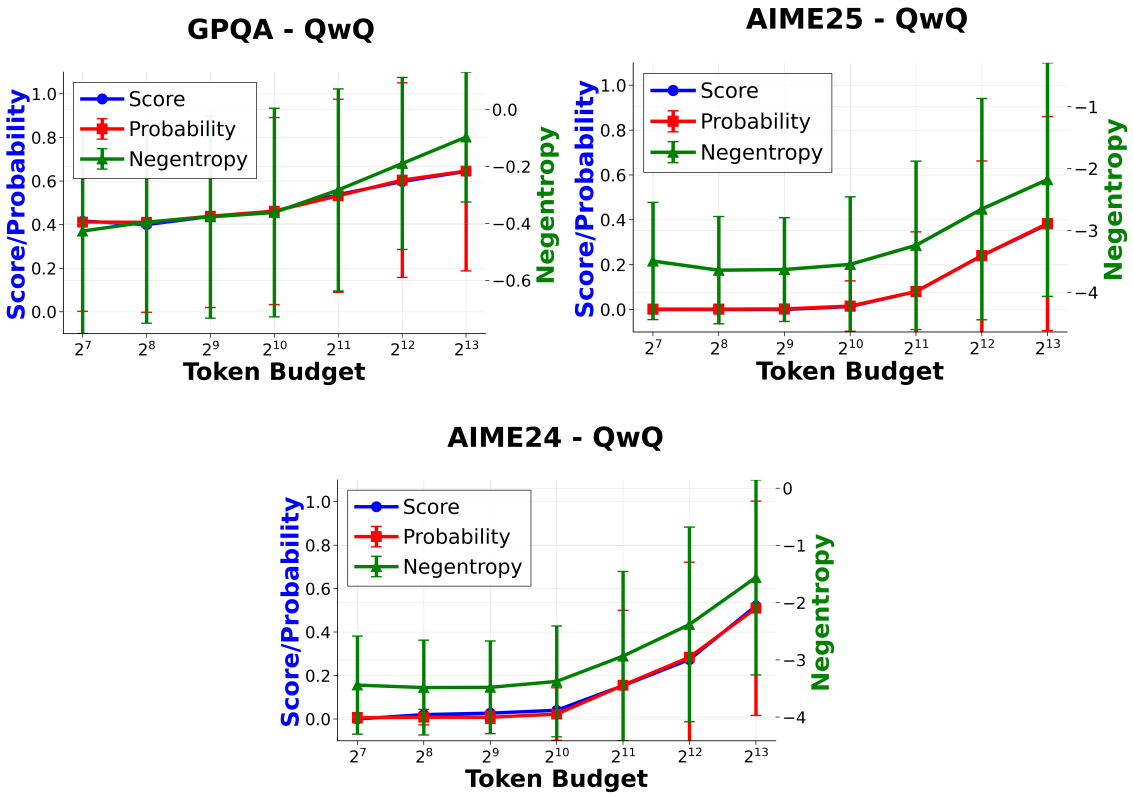
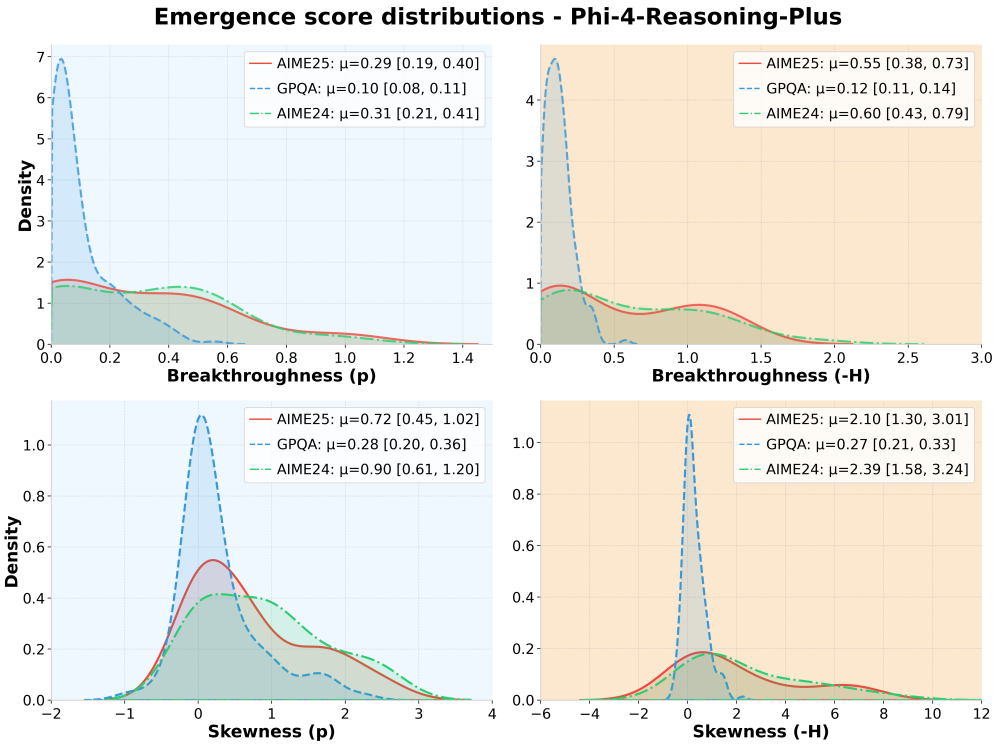


Figure 14: Aggregate scaling curves across datasets for **QwQ-32B**.



	AIME 2024 vs AIME25 $(n_1 = 30, n_2 = 30)$		**AIME 2024 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)	477.00	0.6952	3940.00	0.003981	3663.00	0.0397
Skewness (p)	521.00	0.2973	4067.00	0.0001336	3717.00	0.006256
Breakthroughness (-H)	489.00	0.5692	4795.00	5.99×10^{-8}	4280.00	0.0001004
Skewness (-H)	502.00	0.4464	4799.00	1.13×10^{-9}	4201.00	2.28×10^{-5}

Figure 15: Emergence score distribution across datasets for **Phi-4-Reasoning-Plus**.

Table 5: Results for Mann-Whitney U tests for the emergence score distributions across datasets for **Phi-4-Reasoning-Plus**. Similar to the results for Deepseek-R1-Distill-32B (Section 4.1), nearly all differences between AIME datasets and GPQA are statistically significant, whilst there is no significant difference between AIME24 and AIME25 (at significance level $\alpha = 0.95$).

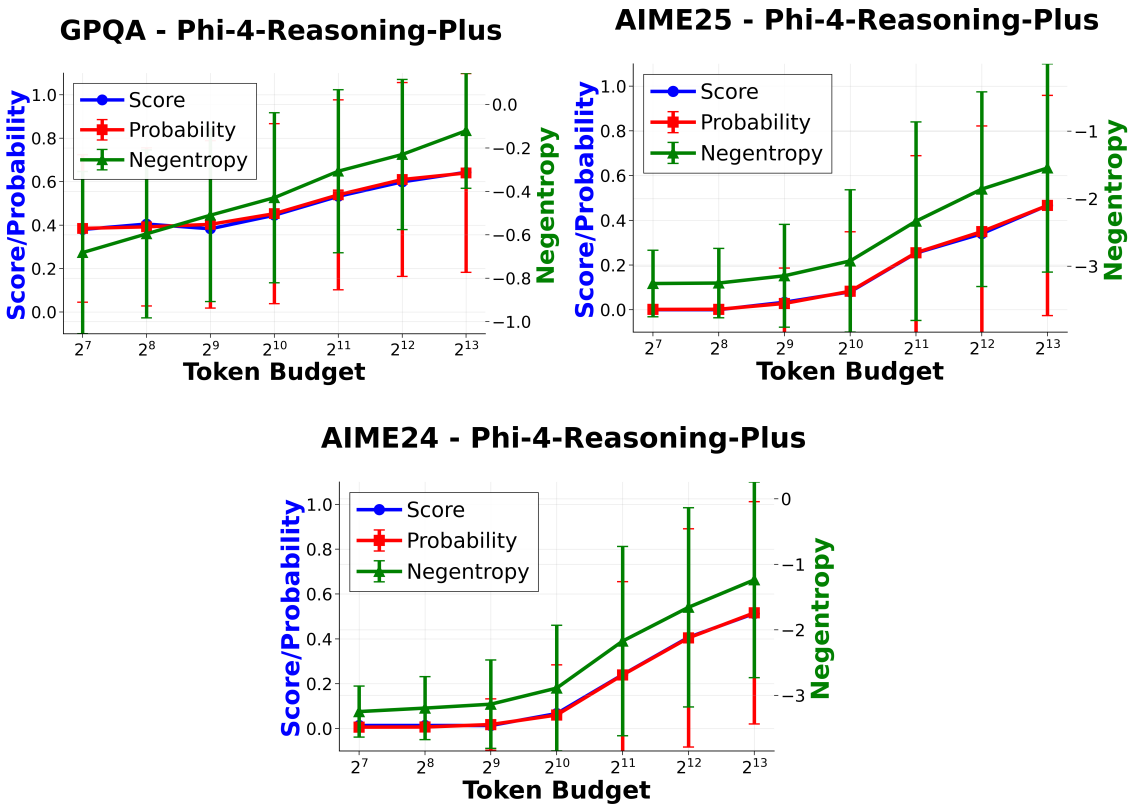


Figure 16: Aggregate scaling curves across datasets for **Phi-4-Reasoning-Plus**.

A.2 Additional top K samples

This appendix shows the top K (K=4) samples for the datasets AIME24 and GPQA, and the corresponding prompts. The main takeaway of Section 4.2 was that emergent scaling trends for individual dataset instances are consistent across models, with the caveat that the onset of abrupt scaling occurs at different token budgets across models. This is also the case for AIME24 and GQPA as shown in Figure 17 and Figure 18. The corresponding prompts are shown in Table 6 and Table 7.

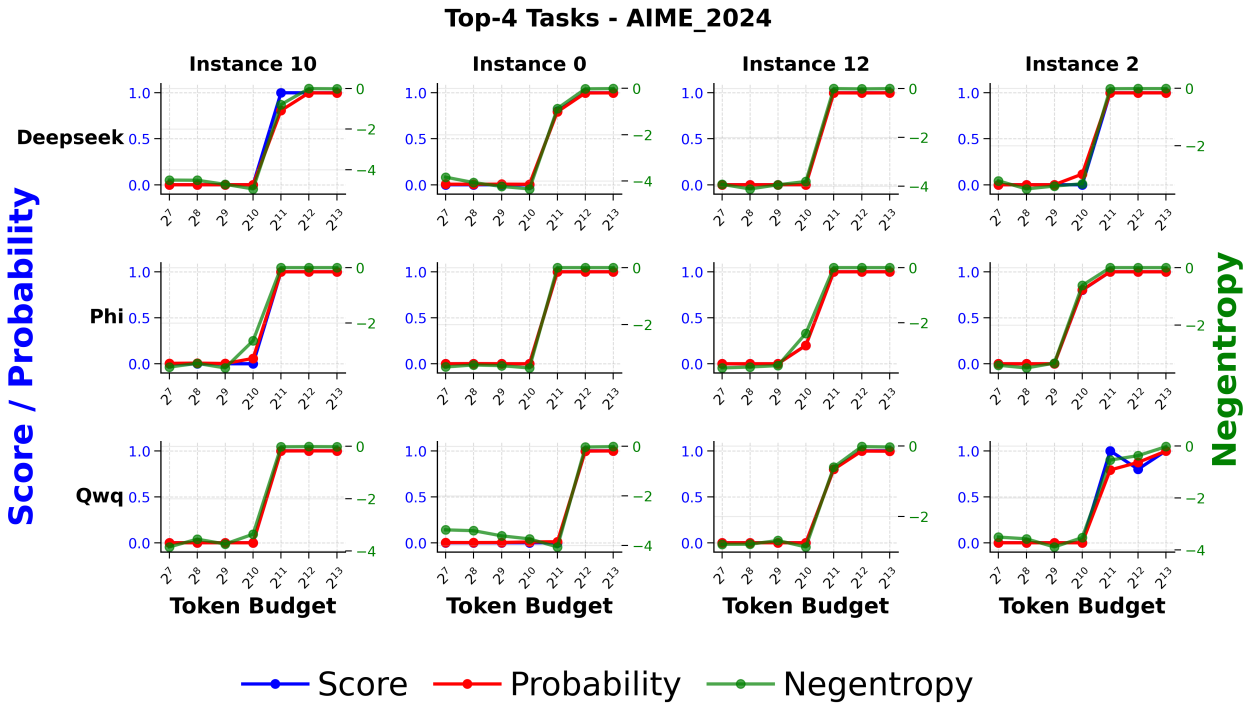


Figure 17: AIME24 top-4 instances displaying emergent test-time scaling - scaling curves

Instance Index	Problem
10	Find the largest possible real part of $$(75+117i)z+\frac{96+144i}{z}$$ where z is a complex number with $\lvert z\rvert = 4$.
0	Let x, y, and z be positive real numbers that satisfy the following system of equations: $$\begin{aligned}\log_2\left(\frac{x}{yz}\right) &= \frac{1}{2}\\ \log_2\left(\frac{y}{xz}\right) &= \frac{1}{3}\\ \log_2\left(\frac{z}{xy}\right) &= \frac{1}{4}\end{aligned}$$ Then the value of $\left\lvert\log_2(x^4y^3z^2)\right\rvert$ is $\frac{m}{n}$ where m and n are relatively prime positive integers. Find $m+n$.
12	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.
2	Jen enters a lottery by picking 4 distinct numbers from $S=\{1,2,3,\ldots,9,10\}$. Four numbers are randomly chosen from S. She wins a prize if at least two of her numbers were among 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 6: AIME24 top-4 instances displaying emergent test-time scaling - prompts

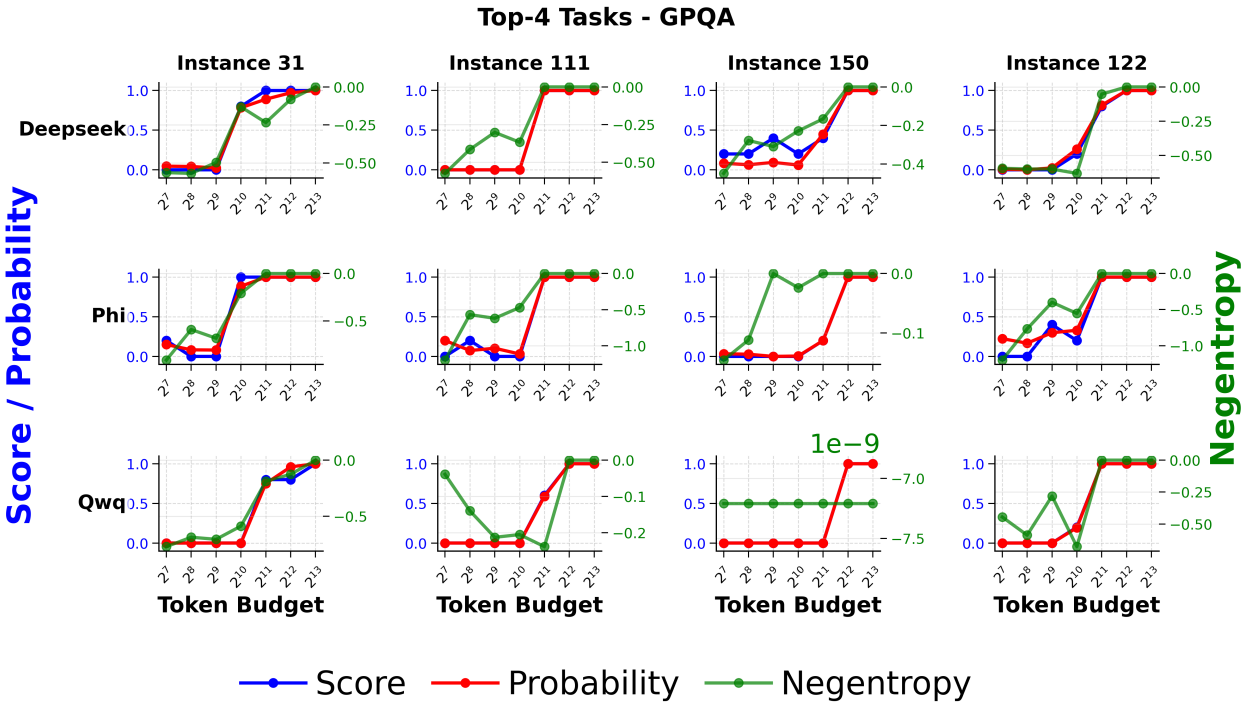


Figure 18: GPQA top-4 instances displaying emergent test-time scaling - scaling curves

Instance Index	Problem
31	What is the energy of the Relativistic Heavy Ion Collider (RHIC) so that the speed of the nucleus X is equal to $0.96c$? Knowing that X is defined as Li with $A = 6$.
111	Let $\|\alpha\rangle$ be the state describing an electron, such that it is proportional to $(1+i)\|\text{up}\rangle + (2-i)\|\text{down}\rangle$, where $\|\text{up}\rangle$ and $\|\text{down}\rangle$ 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.
150	The state of a system at time t is given by the column matrix $\begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}$. An observable of the system is represented by the matrix operator P given by: $$P = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & 0 \end{pmatrix}$$ Calculate the probability that the measurement of the observable will yield 0 at time t.
122	An aspherical supernova type Ic occurred in our galaxy. Part of the supernova ejecta is traveling at a constant velocity of 60 000 km/s (sixty thousand kilometers per second). What distance does the ejecta travel when 50 seconds pass in the ejecta reference frame?

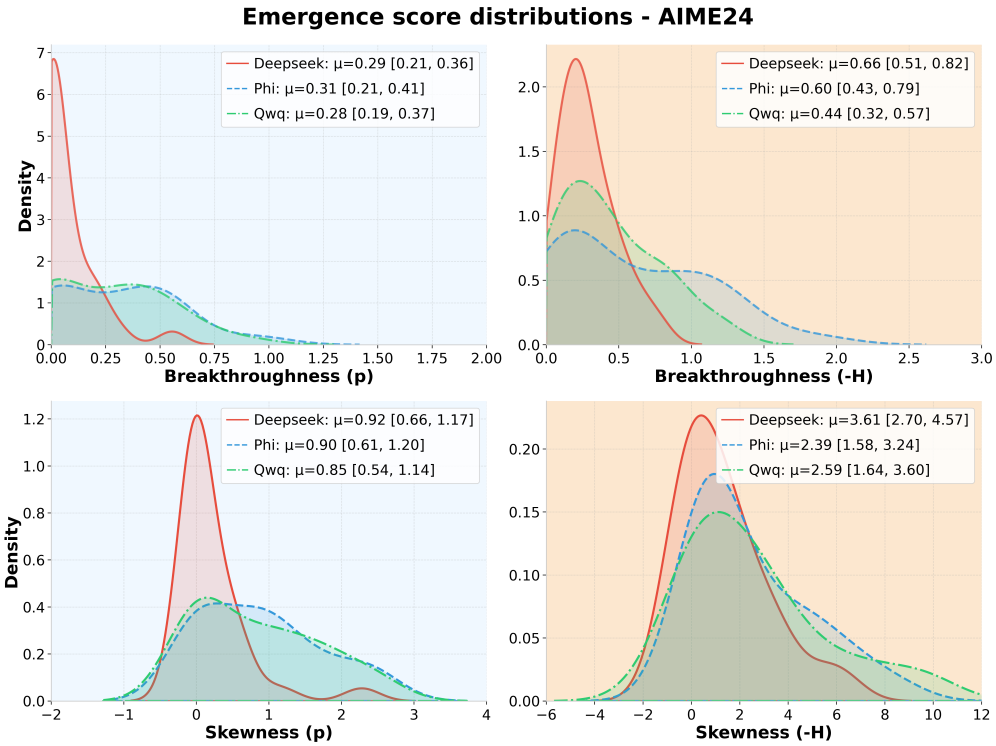
Table 7: GPQA top-4 instances displaying emergent test-time scaling - prompts

A.3 Additional emergence score distributions across models

This section contains the emergence score distributions across models (Deepseek-R1-Distill-32B, QwQ-32B, Phi-4-Reasoning-Plus) when evaluated on AIME24 and GPQA, with the main text discussing results for Deepseek-R1-Distill-32B. The main takeaway of Section 4.3 was that, across models, there is relatively little variation in the sharpness or magnitude of emergent scaling, but the token budget at which abrupt performance changes occured could vary across models. This finding holds almost exactly for AIME24, but is less clear for GPQA.

For AIME24, when testing for statistical significance between the emergence score distributions across model pairs at significance level $\alpha = 0.95$ (Table 8), no difference is found for 11/12 tests, with the one positive result being a borderline case (Breakthroughness (-H) for Deepseek vs QwQ, p=0.02416). When considering the token budget at which emergent scaling occurs (Figure 20), Phi sees slightly earlier onset at 2^9 tokens, whilst the other two models see this begin at 2^{10} tokens.

For GPQA, when testing for statistical significance between the emergence score distributions across model pairs at significance level $\alpha = 0.95$ (Table 9), the picture is more nuanced. There are no statistically significant differences found between the emergence score distributions for the ground truth probability metric, suggesting the sharpness of ground truth probability scaling does not differ across models. This can be observed in the aggregate scaling curves of Figure 22. On the other hand, 4/6 tests comparing negative entropy emergence score distributions are found to be statistically significant, indicating that the degree of abruptness in negative entropy scaling **does** differ across models. However, given that negative entropy scaling in GPQA is substantially smoother than AIME24 and AIME25, with the magnitude of increase in negative entropy being small relative to the other two datasets (this can be seen in Figure 22, as well as by the low mean emergence scores in Figure 21), we do not analyse this particular result further.



	DeepSeek vs Phi $(n_1 = 30, n_2 = 30)$		**DeepSeek vs QwQ** $(n_1 = 30, n_2 = 30)$		**Phi vs QwQ** $(n_1 = 30, n_2 = 30)$	
Emergence Score	U	p	U	p	U	p
Breakthroughness (p)	425.00	0.7172	436.00	0.8418	472.00	0.7506
Skewness (p)	462.00	0.8650	483.00	0.6309	463.00	0.8534
Breakthroughness (-H)	516.00	0.3329	603.00	0.02416	505.00	0.4204
Skewness (-H)	584.00	0.04841	566.00	0.08771	453.00	0.9705

Figure 19: Emergence score distribution across models for **AIME24**.

Table 8: Results of Mann-Whitney U tests for the emergence score distributions across models for **AIME24**. Similar to the results for AIME25 (Section 4.3), nearly all tests find no statistically significant difference, suggesting that the degree of emergent scaling on AIME24 does not differ across models.

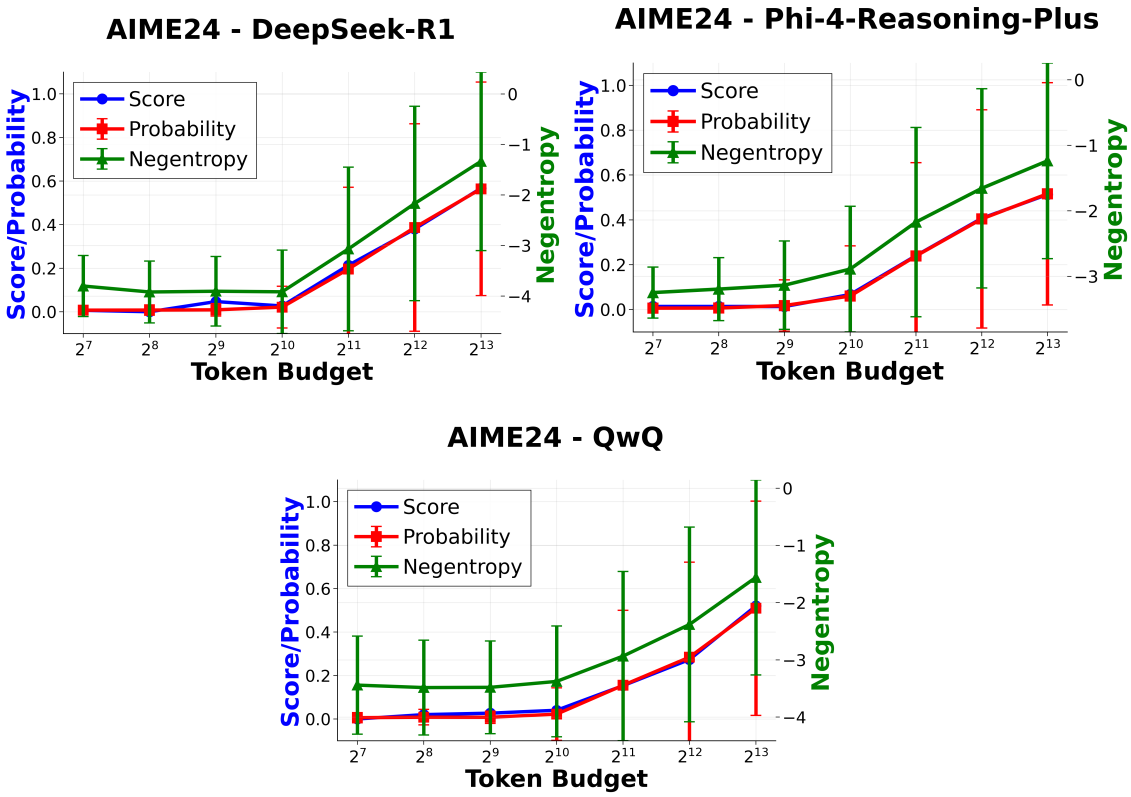
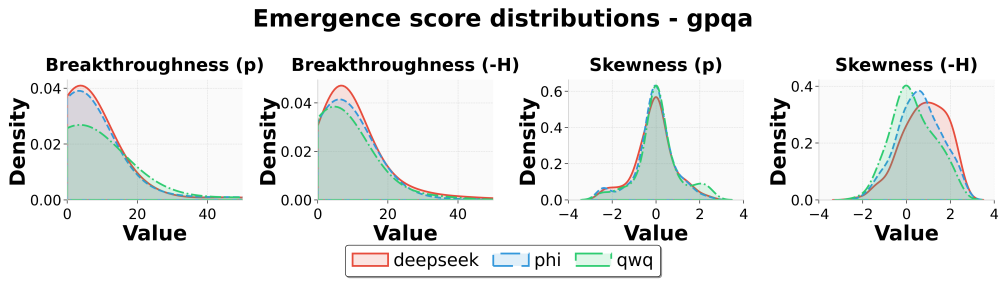


Figure 20: Aggregate scaling behaviour across models for **AIME24**.



	DeepSeek vs Phi ($n_1 = 198, n_2 = 198$)		DeepSeek vs QwQ ($n_1 = 198, n_2 = 198$)		Phi vs QwQ ($n_1 = 198, n_2 = 198$)	
Emergence Score	U	p	U	p	U	p
Breakthroughness (p)	18237.00	0.2308	20189.00	0.6064	21449.00	0.1049
Skewness (p)	18038.00	0.5327	17063.50	0.3775	16999.00	0.7160
Breakthroughness (-H)	19348.00	0.8238	25315.00	5.22×10^{-7}	25829.50	4.52×10^{-8}
Skewness (-H)	20347.00	0.004487	20033.00	6.30×10^{-5}	18145.00	0.1091

Figure 21: Emergence score distribution across models for **GPQA**.

Table 9: Results of Mann-Whitney U tests for the emergence score distributions across models for **GPQA**. No statistical significance is found when comparing emergence distributions for the ground truth probability metric, whilst 4/6 tests comparing negative entropy emergence score distributions are found to be statistically significant. However, the magnitude of increase in negative entropy is small relative to AIME24 and AIME25, so we do not analyse this particular result further.

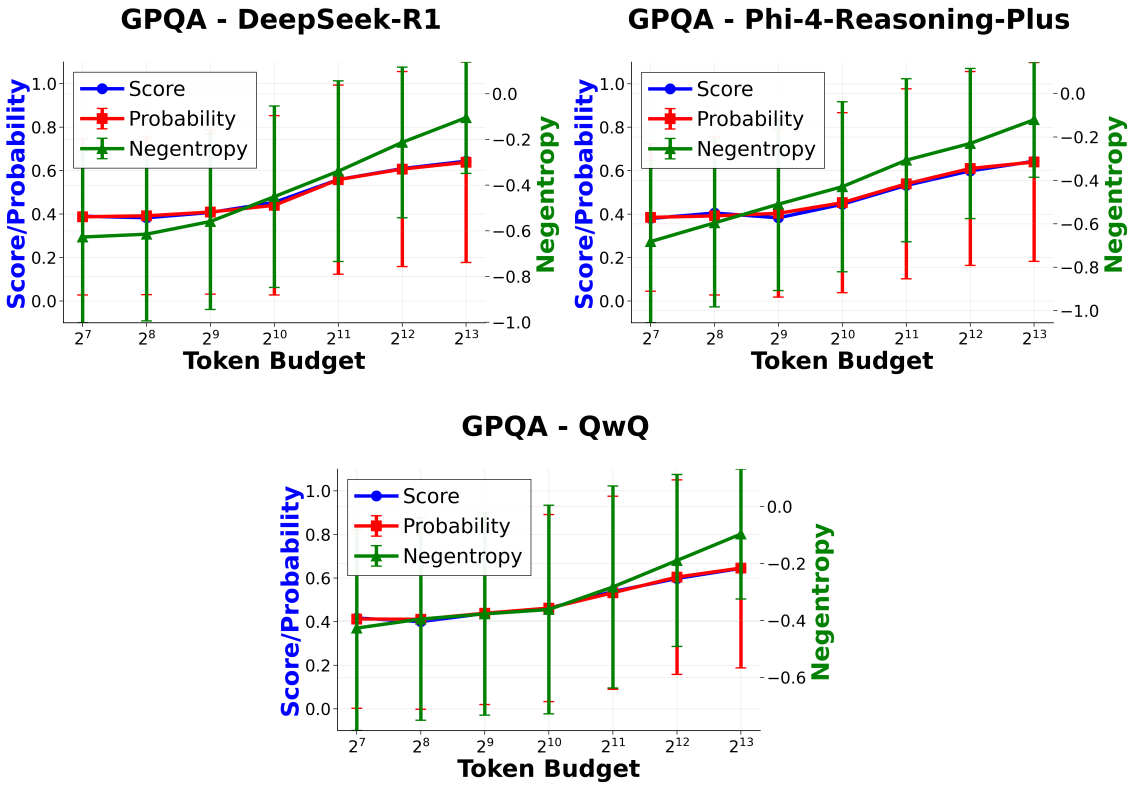


Figure 22: Aggregate scaling behaviour across models for **GPQA**. Note the relatively smooth scaling of nearly all metrics with respect to token budget—negative entropy on Phi-4-Reasoning-Plus is a particularly distinct example. The scaling of negative entropy on QwQ is noticeably sharper, which leads to the positive results for the Mann-Whitney U tests in Table 9. However, the magnitude of increase in negative entropy is small relative to AIME24 and AIME25 (going from ≈ -0.375 bits to ≈ -0.1 bits).

A.4 Additional analysis of emergent scaling across model size

This section presents results for the evaluation of Deepseek-R1-Distill (1.5B, 7B, 14B, 32B) on the AIME24 dataset. A study of GPQA across model sizes was not conducted due to the lack of emergent test-time scaling observed in this dataset. The main takeaway of Section 4.4 was that larger models see sharper scaling of accuracy and ground truth probability, with trends for negative entropy being less consistent—specifically the abruptness of negative entropy scaling across model size was strongly correlated with model size for AIME24, but not for AIME25. The emergence score distributions for AIME24 in Figure 23 support this finding, with sample means, and population mean 95% confidence bounds, monotonically increasing across model size for all distributions. The aggregate scaling curves for AIME24 across model size in Figure 24 further confirm this finding, with the 14B and 32B models displaying more abrupt scaling of ground truth probability **and** negative entropy than the 1.5B and 7B models.

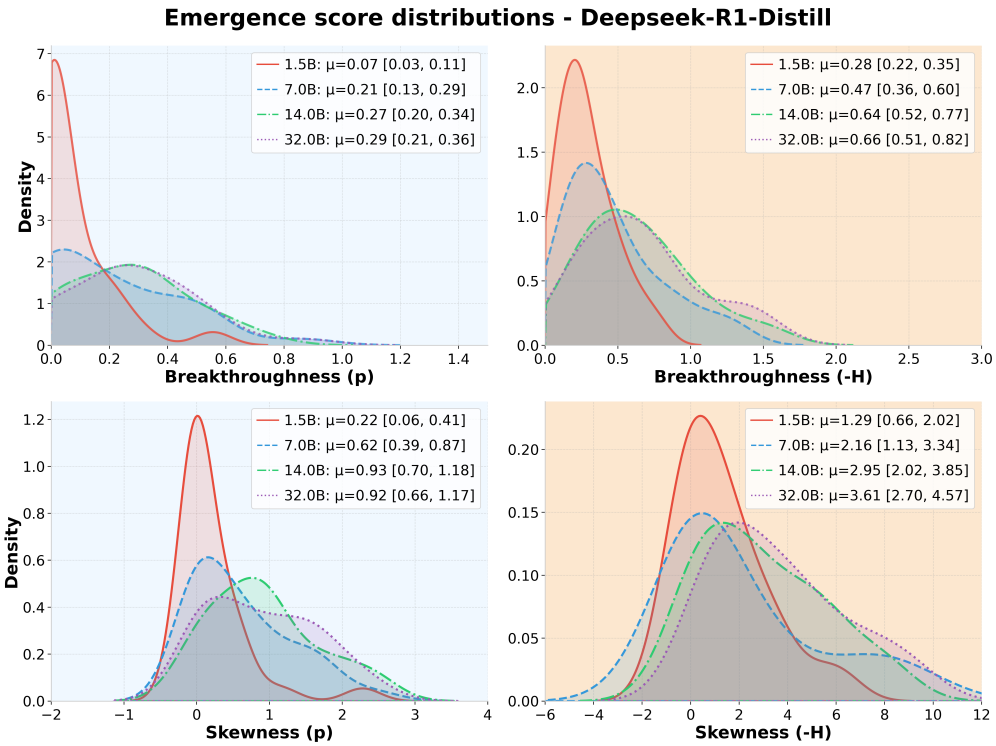


Figure 23: Full distributions of emergence scores across Deepseek model sizes when evaluated on AIME24. Similar to AIME25 (Section 4.4), larger models sees distributions of emergence score shifted rightward for ground truth probability, whilst the smallest model (1.5B) is concentrated around 0. *Unlike AIME25*, the emergence score distributions for the negentropy metric are also shifted rightward for larger models, suggesting that abrupt increases in model confidence are also more likely to occur for larger models.

The top four instances in AIME24 displaying emergent scaling are shown in Figure 25—similar to the aggregate scaling curves above, the larger models show noticeably sharper scaling of all three tracked metrics.

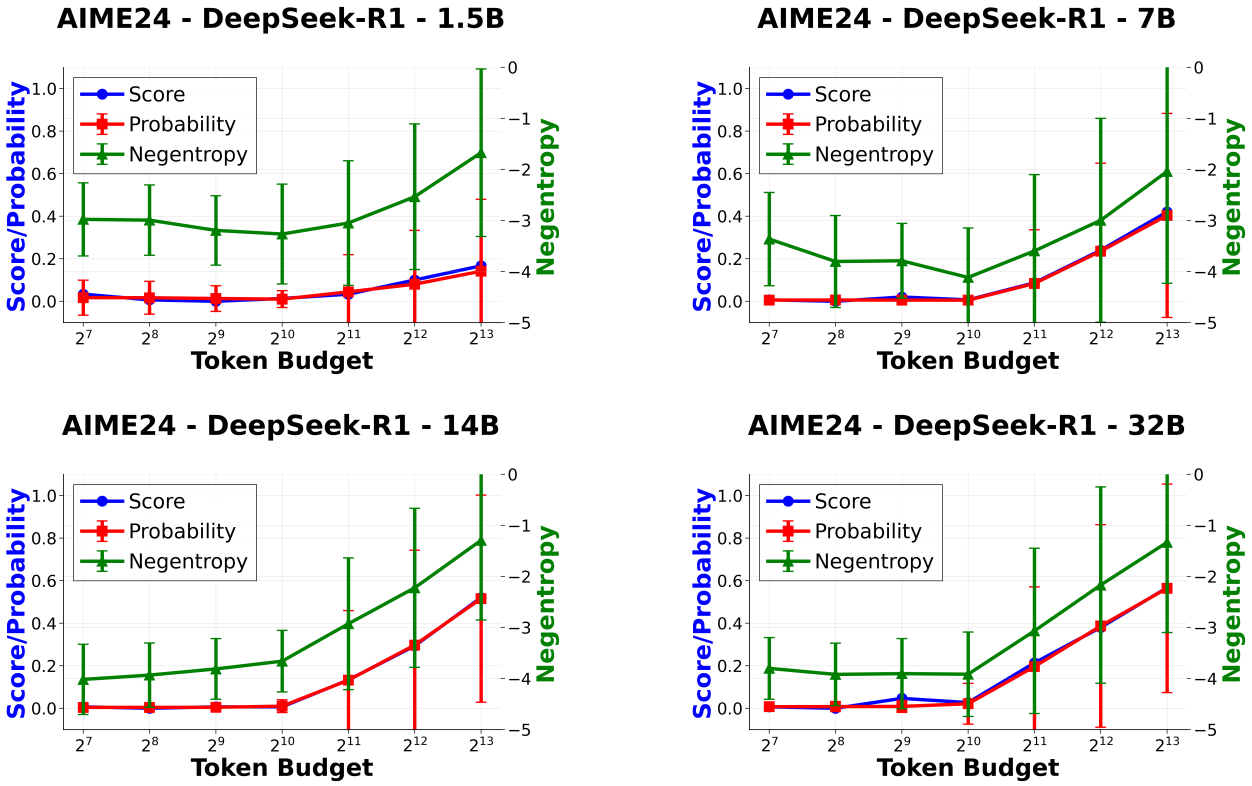


Figure 24: Aggregate scaling behaviour across Deepseek model sizes when evaluated on AIME24. 1.5B exhibits very smooth scaling, where 7B remains at 0 accuracy/ground truth probability until 2^{10} tokens before increasing to 40% accuracy at 2^{13} tokens. Both 14B and 32B also remain at 0 accuracy and ground truth probability until 2^{10} tokens but increase more rapidly to final accuracies at 2^{13} tokens, with 32B attaining 60%. *Unlike AIME25*, there is a stronger correlation between accuracy/ground truth probability and negentropy, with the better performing models (14B and 32B) achieving lower final negentropy than the 1.5B and 7B models.

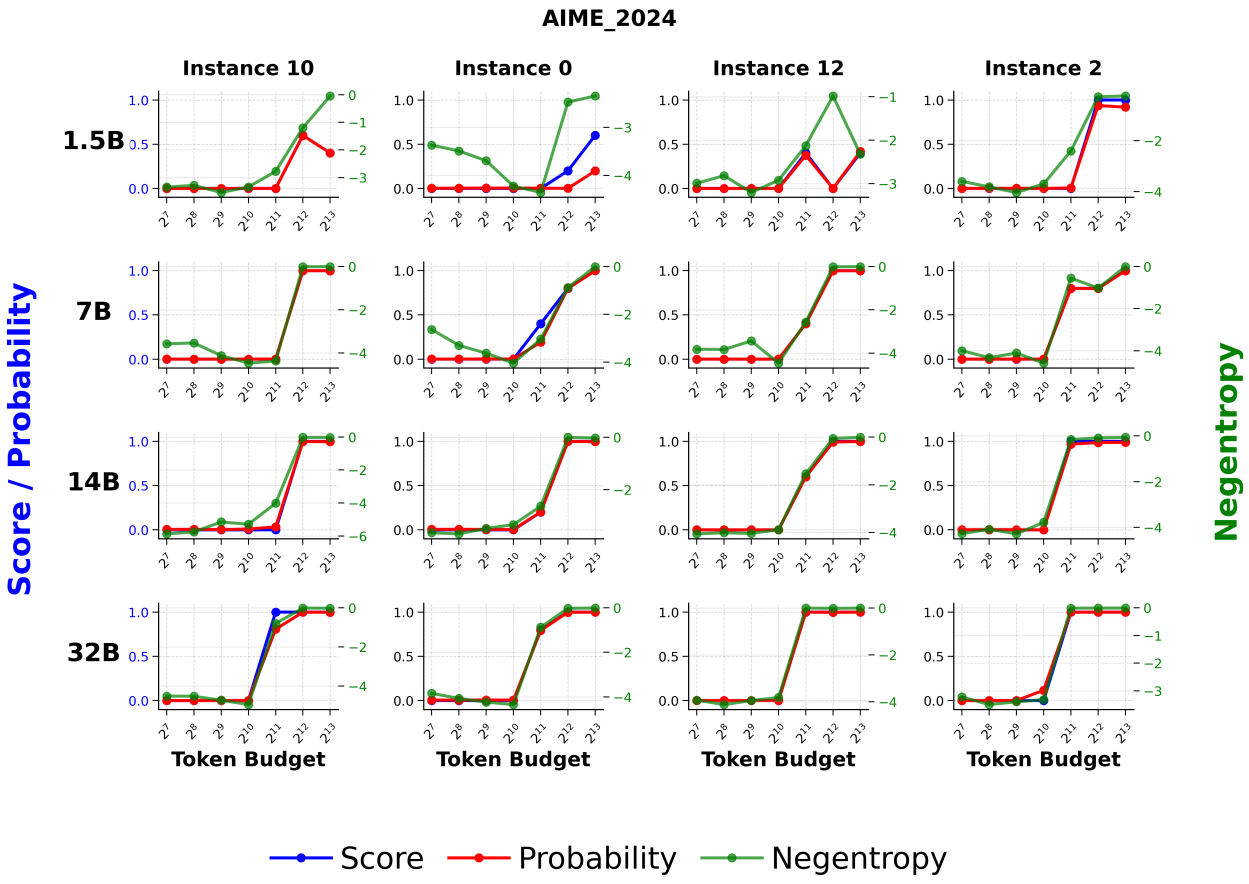


Figure 25: Top-k AIME2024 instances exhibiting emergent test-time scaling across model sizes.