Emergent Abilities during Inference Time Scaling

Iyngkarran Kumar Edoardo M. Ponti
i.kumaraguruparan@sms.ed.ac.uk
University of Edinburgh

Abstract

This study investigates the existence of emergent scaling behaviour in language reasoning models as test-time compute budget is scaled. We study scaling behaviour of three popular open-weight reasoning models on GPQA-Diamond, AIME2025, and AIME2024, and examine the frequency of sharp, non-linear increases in performance as test-time compute budget is scaled. We also conduct a study across Deepseek-R1-Distill model size, examining the influence of parameter count on the tendency for test-time scaling behaviour to exhibit emergent properties. In doing so, we investigate the predictability of the test-time scaling behaviour of reasoning models, and whether certain abilities may spontaneously emerge at given inference compute budgets.[1]

1 Introduction

Large language models have shown remarkable scaling behavior with respect to model parameters, dataset size, and training compute (Kaplan et al., 2020). These "scaling laws" allow some abilities of frontier models to be predicted by training counterparts that take a fraction of the resources to create (Achiam et al., 2023). This consistent scaling behavior has both 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 language models — specific tasks for which models exhibit sharp, non-linear improvement when scaled (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.

Questions about the predictability of scaling behaviour are relevant in test-time settings too, with previous studies establishing "test-time compute scaling laws" (Snell et al., 2024; Muennighoff et al., 2025) - linear relationships between the logarithm of test-time compute budget and model performance. However, prior work has only examined the scaling behaviour of models on aggregate benchmarks. When evaluating models on more granular task distributions, emergent scaling behaviour may be more likely to occur. This study is concerned with the continuity and predictability of test-time compute scaling; are sudden breakthrough moments in reasoning traces (such as the "Aha!" moments in (Guo et al., 2025)) common or rare occurrences in language reasoning models (LRMs), and are they fundamental features of the models and tasks themselves, or artifacts that arise from experimental design (Schaeffer et al., 2023)?

1.1 Research questions

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

1. To what extent are sharp, non-linear increases in performance (emergent scaling behaviour) observed as test-time compute is scaled? How is test-time emergent scaling influenced by:
   (a) The task distribution?
   (b) The reasoning trajectory?
   (c) The reasoning model used for evaluation (model size and model type)?

[1] Code for this project (in progress) is available at https://github.com/IyngkarranKumar/test_time_emergent_scaling

2. Is observed emergent test-time scaling contingent on using discrete metrics to evaluate performance ((Schaeffer et al., 2023)), or are they present across both discrete and continuous scoring metrics?

2 Related Work

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

2.1 Emergent ability of large language models

2.1.1 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). Note that 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). A consistent definition of emergent scaling that is used in this work is a sharp, non-linear increase 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 computation, such as few-shot 3 digit addition and French-English translation. Additionally, the majority of these tasks were evaluted 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, the GPT-3 paper 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%. A list of 137 emergent abilities is given in (Wei, 2022).

2.2 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 by 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, that 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 abilties is required; follow-up work has questioned the appropriateness of proposed continuous metrics such as Token Edit Distance for evaluating model performance (Berti et al., 2025), and identified some methodological issues in the definition of emergent scaling used in (Schaeffer et al., 2023), arguing that this work definies emergent scaling too strictly (Berti et al., 2025). Additionally, many other works identify tasks that exhibit sharp, nonlinear 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.

2.2.1 Mechanisms underlying emergent scaling behavior

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 domain (Okawa et al., 2024) and natural language domain (Lubana et al., 2024). (Okawa et al., 2024) study how a diffusion model learns how to independently generate shapes of distinct colour and shape, before abruptly learning how to compose these concepts to generate out-of-distribution shapes. (Lubana et al., 2024) also conceptualize 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 generalization). (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 memorization and generalization circuits in neural models, again, reinforcing the point that training dynamics have a strong influence on the likehlihood 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 categorize tasks into those that exhibit U-shaped and inverted U-shaped scaling, and show that aggregating these curves can produce the sudden "breakthrough" in model performance that characterises emergent scaling. Their hypothesed explanation for emergent scaling is intuitive, but was only tested on multiple-choice tasks and remains to be extended to other settings. In summary, there may be many different underlying causes of emergent scaling in language models, some of which are listed above, and some which remain to be discovered. Note that none of the proposed causes above are mutually exclusive, or is broadly accepted as the primary mechanism that leads to emergent scaling; thus, this topic is still an active area of research.

2.3 Scaling test-time compute in language reasoning models

2.3.1 Using compute at test-time

There are a wide range of methods for leveraging compute at test-time to enhance model responses, broadly clustered into sequential and parallel methods. Sequential methods increase the length of the reasoning trace that a model produces, such as by training 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). Doing so leads to a refined output distribution relative to the base model (Snell et al., 2024). Finetuning base-models to update their initial proposal distribution in these ways has been incredibly effective in improving model performance on complex reasoning tasks, such as mathematical and scientific benchmarks (Guo et al., 2025; Singh et al., 2023; Zelikman et al., 2022).

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). A score function may be applied to the solutions to determine the best response (Cobbe et al., 2021), or to intermediate steps taken during reasoning (Lightman et al., 2023), with methods such as beam search (Feng et al., 2024), lookahead search, or Monte Carlo Tree Search (MCTS) (Sutton et al., 1999) then used to prune the search space.

2.3.2 Scaling laws for test-time compute

The proliferation of methods to leverage test-time compute has motivated a body of 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 begin 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 (Gema et al., 2025). Relatedly, (Marjanović et al., 2025) finds that U-shaped scaling behaviour can be observed as reasoning trace length is scaled, with minimal degradation in model performance when reducing the trace length by up to 50%. (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, avoiding the computational cost of methods such as MCTS that results from explicitly performing rollouts.

To situate the current study within the literature, a broad body of work has studied the relationship between test-time compute and model performance in aggregate. However, few have explored the continuity and predictability of test-time scaling - specifically, to what extent does sharp, non-linear (emergent) scaling occur when test-time compute is scaled, and what factors influence this? Through a qualitative lens, this study is concerned with cases in which models make sudden progress towards the correct answer when reasoning - in other words, when do language models exhibit "Aha!" moments (Guo et al., 2025), and do these represent a breakthrough moment in model reasoning?

3 Method

3.1 Scaling test-time compute

Broadly there are two ways to scale test-time compute, sequential scaling and parallel scaling. Sequential scaling increases the length of the reasoning trace that a model produces when responding to a prompt. Parallel scaling leverages additional compute to sample multiple outputs from the model, then aggregates these outputs using a method such as majority voting (Wang et al., 2022), or best-of-N sampling (Cobbe et al., 2021; Lightman et al., 2023) . Whilst a combination of these methods can be used to scale test-time compute, this study focuses on sequential scaling due to the more natural interpretation of emergent scaling behaviour as a function of reasoning trace length[2].

To control the length of a reasoning trace, we use the *budget forcing* method of (Muennighoff et al., 2025), and append special tokens onto the end of a reasoning trace to force a model to end, or continue, it's output. Specifically, to force a model to output a reasoning trace and answer of length T tokens (the token budget), we sample from the model in standard fashion; however, if the model tries to end its response before hitting the token budget (by returning an `<EOS>` token), we append the sequence `Hmm, let me keep thinking` to force the model to reconsider it's answer and continue reasoning. 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 $N_f = 20$ 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 LRMs on these tasks (Xu et al., 2025). Specifically, the following benchmarks are used:

1. American Invitational Mathematics Examination 2024 and 2025 (AIME-2024, AIME-2025). This datasets contains 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.

We study test-time scaling behaviour on three medium-sized open-weight reasoning models:

- DeepSeek-R1-Distill-32B (Guo et al., 2025):
- QwQ-32B (QwenTeam, 2025)

[2] Our hypothesis is that a reasoning model may make sudden progress towards a solution as it reasons for **longer**; on the other hand, we find it unlikely that sampling more responses from a model will lead to a sudden breakthrough in performance.

[3] This is a parameter that can be varied in our setup.

- Phi-4-Reasoning-Plus (14B) (Abdin et al., 2025)

Additionally, emergent test-time scaling is studied with respect to model size by evaluating against the Deepseek-R1-Distill family - specifically, for the 1.5B, 7B, 14B, and 32B models (Guo et al., 2025).

3.3 Metrics

At each token budget, a reasoning model M outputs a reasoning trace r and answer y^* to input prompt x, which is assessed against the ground truth answer y.

Note that the benchmarks in section 3.2 have finite 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 over it's vocabulary can be renormalized over the solution set $\mathcal{S}$.

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, renormalized 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 renormalized 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 renormalized probability of solution s.

Notice that 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 progress measures.

3.4 Quantitative proxies for emergence

We quantify emergent scaling behavior across test-time compute budgets using two complementary metrics:

1. **Breakthroughness Score** (Srivastava et al., 2022)

2. **Difference Distribution Skewness** (ours)

Breakthroughness Score: For performance data $\{(x_i, y_i)\}_{i=1}^n$ ordered by compute budget x_i, the breakthroughness score is:

$$B = \frac{I(y)}{\text{RootMedianSquare}(\{y_{i+1} - y_i\}_i)} \tag{1}$$

where $I(y) = \text{sign}(\arg\max_i y_i - \arg\min_i y_i) \cdot (\max_i y_i - \min_i y_i)$ captures the signed magnitude of the overall performance change.

Difference Distribution Skewness: Due to limitations of the breakthroughness metric (discussed in Appendix A), we use a complementary measure based on the Fisher-Pearson skewness coefficient of performance differences (also referred to as performance deltas). Let $\Delta y_i = y_{i+1} - y_i$ denote consecutive performance deltas. The skewness score is:

$$S = \frac{\sqrt{N(N-1)}}{N-2} \cdot \frac{\frac{1}{N}\sum_{i=1}^{N-1}(\Delta y_i - \overline{\Delta y})^3}{s_\Delta^3} \tag{2}$$

where $\overline{\Delta y}$ is the mean difference, s_Δ is the standard deviation of differences, and $N = n - 1$ is the number of differences. Large positive skewness indicates a performance delta distribution that has probability mass concentrated at lower values with a long right tail, which is characteristic of an emergent scaling curve. See section A for further discussion.

4 Results

This section is structured as follows. First aggregate scaling behaviour and correlation coefficients between discrete and continuous metrics are presented (section 4.1). Then, trends in the emergent score distributions across datasets (section 4.2), models (section 4.3), and model scales (section 4.4) are examined. Finally, task instances with high emergence scores are studied, and the dependence of significant test-time emergent scaling behaviour on the task instance and reasoning trajectory is investigated (section 4.5).

4.1 Aggregate scaling behaviour across datasets

See Figure 1 for aggregate scaling curves, and Figure 2 for Pearson correlations between discrete and continuous metrics.

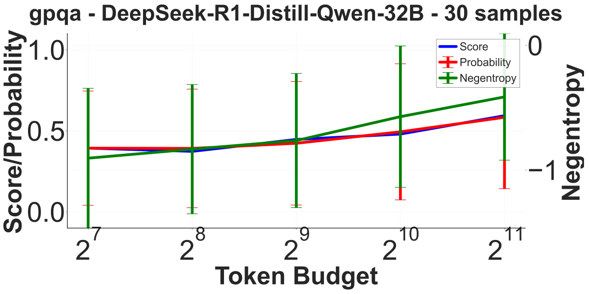

Figure 1: Scaling curves for score, probability of ground truth, and negative entropy over the solution set. Results presented for Deepseek-R1-Distill-32B on GPQA (30 samples).

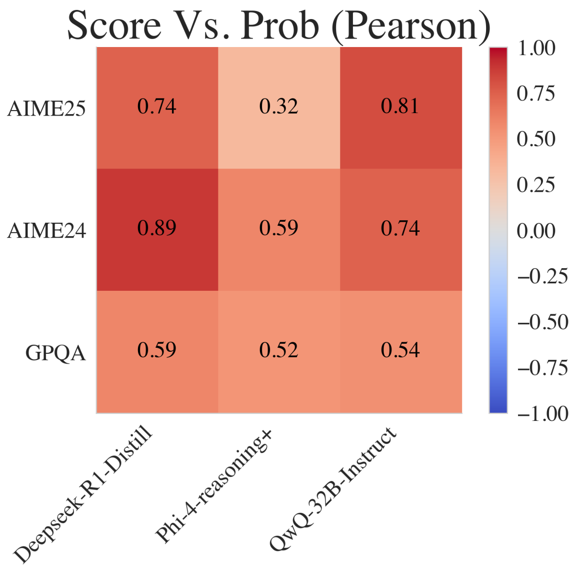

Figure 2: Pearson correlations between accuracy and ground truth probability (averaged over all samples) for all model-dataset pairings. As expected, moderate-to-strong correlations between discrete and continuous metrics are observed.

4.2 Emergent scaling trends across datasets

See Table 1 for summary statistics for the emergence scores distributions across datasets, and Figure 3 for full distributions.

Dataset	B_p	B_{-H}	S_p	S_{-H}
GPQA	(3.09, 2.85)	(3.6, 2.9)	(0.372, 1.09)	(0.456, 1.04)
AIME25	(2.53, 2.5)	(2.96, 4.91)	(0.717, 1.16)	(0.709, 1.04)
AIME24	(2.96, 2.16)	(3.01, 2.48)	(0.51, 1.2)	(0.456, 1.07)

Table 1: Summary statistics (center, spread) for the emergence scores distributions across GPQA, AIME25, and AIME24 using Deepseek-R1-Distill-32B. Median and interquartile range are presented for the breakthroughness metrics, whereas mean and standard deviation are presented for the skewness metrics, due to the tendency for the breakthroughness scores to spread over two orders of magnitude. Emergence scores are referenced as follows: B_p is breakthroughness score for probability of ground truth, B_{-H} is breakthroughness score for negative entropy, S_p is skewness score for probability of ground truth, S_{-H} is skewness score for negative entropy. GPQA has the highest average breakthroughness scores across both probability and entropy metrics, while AIME25 shows the highest skewness values, suggesting these datasets exhibit different patterns of test-time emergent scaling behaviour that are captured by these two metrics.

4.3 Emergent scaling trends across models

See Table 2 for summary statistics for the emergence scores distributions across models, and Figure 4 for full distributions.

Model	B_p	B_{-H}	S_p	S_{-H}
Deepseek-R1	(3.09, 2.85)	(3.6, 2.9)	(0.372, 1.09)	(0.456, 1.04)
QwQ	(2.99, 2.32)	(2.63, 1.72)	(0.7, 0.85)	(-0.0419, 1.13)
Phi-4-reasoning-plus	(2.8, 2.0)	(2.86, 1.87)	(0.129, 1.1)	(0.371, 1.14)

Table 2: Summary statistics for emergence scores across Deepseek-R1, QwQ, and Phi-4-reasoning-plus, when evaluated on GPQA. Observe that Deepseek-R1 has the highest average scores across 3 out of 4 metrics, indicating that this model is most likely to display test-time emergent scaling behaviour.

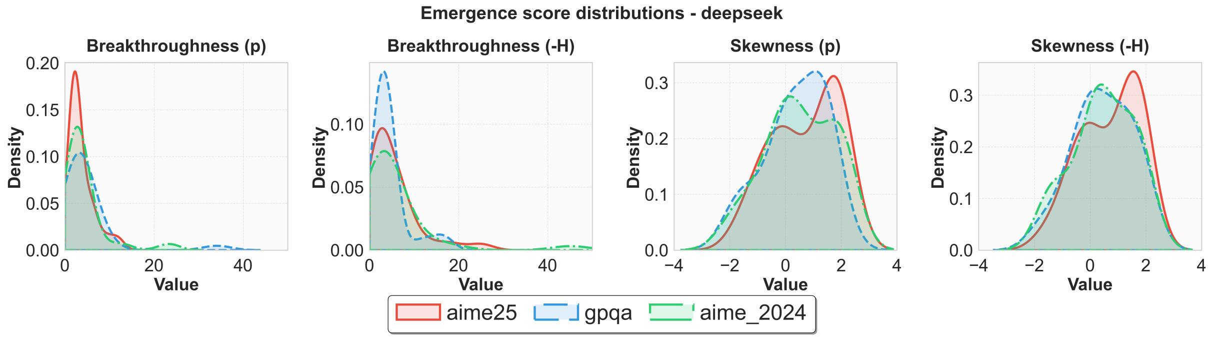

Figure 3: Full distributions of emergence scores across datasets when evaluated on Deepseek-R1-Distill-32B. Summary statistics are shown in Table 1. Distributions with modes shifted towards higher values indicate a greater degree of emergent test-time scaling.

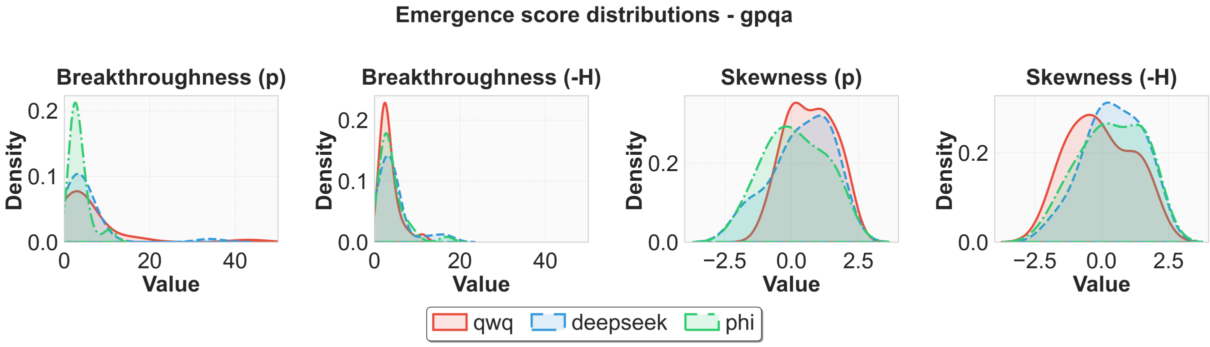

Figure 4: Full distributions of emergence scores across models when evaluated on GPQA. Summary statistics are shown in Table 2. Distributions with modes shifted towards higher values indicate a greater degree of emergent test-time scaling.

4.4 Emergent scaling trends across model scales

See Figure 5 for distribution centre and spread statistics for the emergence scores across Deepseek-R1-Distill model size, and Figure 6 for full distributions.

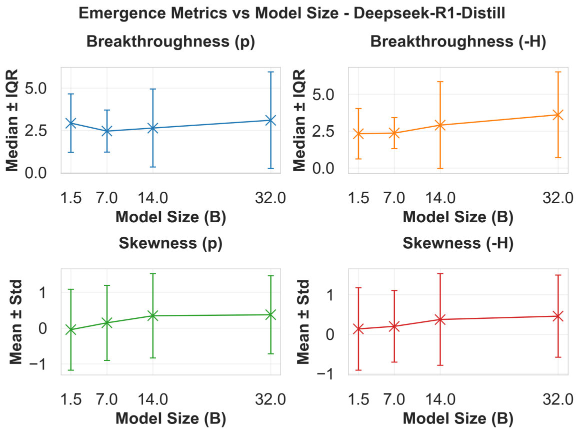

Figure 5: Distribution centre and spread statistics for the emergence scores (median and interquartile range for the breakthroughness metrics, mean and standard deviation for the skewness metrics) across Deepseek-R1-Distill model size. There is a general tendency for emergence scores to increase with model scale, suggesting that larger models are more likely to display test-time emergent scaling behaviour.

4.5 Task instances with high emergence scores

See Figure 7 for samples from GPQA with highest emergence scores when evaluated with Deepseek-R1-Distill-32B.

Further analysis will be performed to investigate the dependence of test-time emergent scaling behaviour on the task instance and reasoning trajectory.

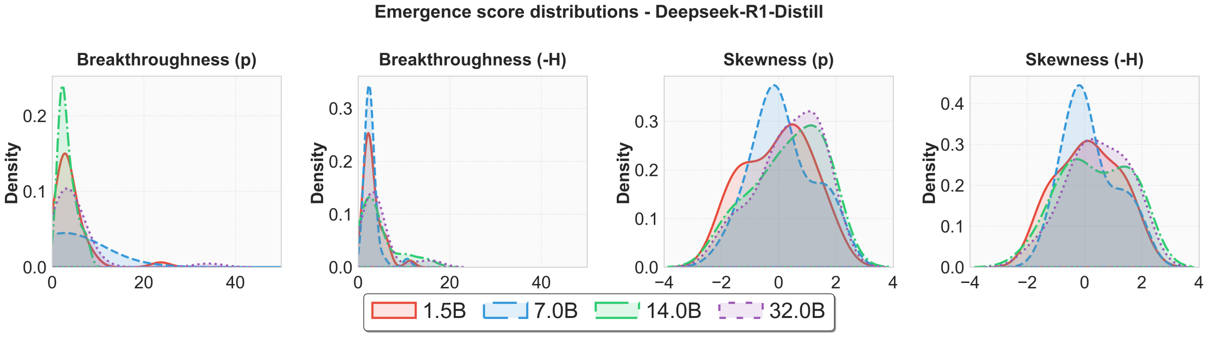

Figure 6: Distributions of emergence scores across Deepseek-R1-Distill model size. Summary statistics are shown in Figure 5. Distributions with modes shifted towards higher values indicate a greater degree of emergent test-time scaling.

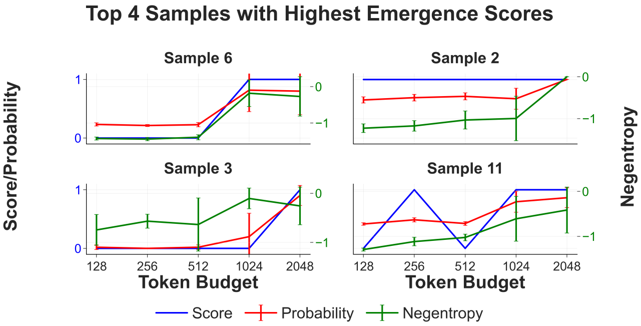

Figure 7: Samples from GPQA with highest emergence scores (when tracking the probability metric) when evaluated with Deepseek-R1-Distill-32B. Notice the non-linear scaling of ground truth probability with respect to token budget, which deviate from smooth, predictable scaling behaviour. This is especially pronounced for task instances 6 and 3.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. 2025. Phi-4-reasoning technical report.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Sanjeev Arora and Anirudh Goyal. 2023. A theory for emergence of complex skills in language models.

Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. 2025. Emergent abilities in large language models: A survey.

Bradley Brown, Jordan Juravsky, Ryan Saul Ehrlich, Ronald Clark, Quoc V Le, Christopher Re, and Azalia Mirhoseini. 2025. Large language monkeys: Scaling inference compute with repeated sampling.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. 2025. Understanding emergent abilities of language models from the loss perspective.

Xidong Feng, Ziyu Wan, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2024. Alphazero-like tree-search can guide large language model decoding and training.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, 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.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, et al. 2025. Inverse scaling in test-time compute. *arXiv preprint arXiv:2507.14417*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Shengding Hu, Xin Liu, Xu Han, Xinrong Zhang, Chaoqun He, Weilin Zhao, Yankai Lin, Ning Ding, Zebin Ou, Guoyang Zeng, et al. 2023. Predicting emergent abilities with infinite resolution evaluation. *arXiv preprint arXiv:2310.03262*.

Yufei Huang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Unified view of grokking, double descent and emergent abilities: A perspective from circuits competition. *arXiv preprint arXiv:2402.15175*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Ekdeep Singh Lubana, Kyogo Kawaguchi, Robert P. Dick, and Hidenori Tanaka. 2024. A percolation model of emergence: Analyzing transformers trained on a formal language.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback, 2023. *URL https://arxiv. org/abs/2303.17651*.

Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, et al. 2025. Deepseek-r1 thoughtology: Let's think about llm reasoning. *arXiv preprint arXiv:2504.07128*.

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

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2021. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. 2024. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

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

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

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

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

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. 2023. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

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

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

Liming Wang, Siyuan Feng, Mark Hasegawa-Johnson, and Chang Yoo. 2022. Self-supervised semantic-driven phoneme discovery for zero-resource speech recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8027–8047, Dublin, Ireland. Association for Computational Linguistics.

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

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Tung-Yu Wu and Pei-Yu Lo. 2024. U-shaped and inverted-u scaling behind emergent abilities of large language models. *arXiv preprint arXiv:2410.01692*.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning.

Appendix

A Emergence score metrics

This study uses two metrics to quantify the degree of emergent scaling behaviour: Breakthroughness score (Srivastava et al., 2022) and the skewness of the difference distribution. This section provides more details on these metrics, such as the different types of scaling behaviour they capture.

First, recall that these metrics are defined as follows:

Breakthroughness Score:

$$B = \frac{I(y)}{\text{RootMedianSquare}(\{y_{i+1} - y_i\}_i)} \tag{3}$$

Where $I(y)$ is the signed magnitude of the overall performance change: $I(y) = \text{sign}(\arg\max_i y_i - \arg\min_i y_i) \cdot (\max_i y_i - \min_i y_i)$. In this study we are interested in the magnitude of the change so the absolute value of $I(y)$ is used.

Difference Distribution Skewness:

$$S = \frac{\sqrt{N(N-1)}}{N-2} \cdot \frac{\frac{1}{N}\sum_{i=1}^{N-1}(\Delta y_i - \overline{\Delta y})^3}{s_\Delta^3} \tag{4}$$

where $\Delta y_i = y_{i+1} - y_i$, $\overline{\Delta y}$ is the mean difference, s_Δ is the standard deviation of differences, and $N = n - 1$ is the number of differences.

Table 1 showed that GPQA has the highest average scores across the breakthroughness metrics, whilst AIME25 has the highest skewness values. One interpretation of this result is that these datasets display different patterns of emergent scaling.

The Breakthroughness metric has been used in previous studies of emergent abilties (Srivastava et al., 2022; Schaeffer et al., 2023), however the two properties listed below make the metric less than ideal for capturing emergent scaling.

1. **Sensitivity to large range of response variable (y)**: The numerator $I(y)$ takes large values when the performance scores span a wide range, meaning that rapidly growing but continous functions, such as an exponential, will be assigned high breakthroughness scores. This can be seen in Table 3 - in the domain $x \in [0, 1]$ exponential functions are approximately linear, meaning that an ideal emergence score metric would be low. However, the exponential functions are assigned high breakthroughness scores.

2. **Invariance to qualitatively different emergent scaling curves**: Consider the one-step and two-step functions given in Table 3. Both functions have abrupt, but qualitatively different scaling behaviour. However, the breakthroughness metric treats these curves equivalently, due to its consideration of only the difference between the maxium and minumum values of the response variable, and the average (root median square) of the differences. However, the skewness metric correctly distinguishes between the two curves, assigning a higher score to the one-step function.

Function	**Breakthrough**	**Skewness**
e^x	24.51	-0.05
$e^{0.5x}$	15.35	-0.06
$e^{2(x-0.5)}$	62.96	-0.21
$\mathbb{1}_{x \geq 0.5}$ (one-step)	27.92	4.33
$0.5 \cdot \mathbb{1}_{x \geq 0.5} + 0.5 \cdot \mathbb{1}_{x \geq 0.7}$ (two-step)	27.06	0.97

Table 3: Breakthroughness and skewness scores for different functions. The breakthroughness metrics has two properties that make it suboptimal as a proxy for emergent scaling behaviour: (a) assigning high scores to smooth functions that span a large range (e.g. exponential functions), and (b) treating qualitatively different emergent scaling curves equivalently (e.g. one-step and two-step functions).