Emergent Abilities during Inference Time Scaling

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

DRAFT FOR FELLOWSHIP APPLICATIONS - Work in progress. Please do not share. See footnote for further detail.[1]

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 capabilities may spontaneously emerge at given inference compute budgets.[2]

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 capabilities 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 capabilities and potential risks without needing to commit significant resources for training.

However, a number of studies have identified "emergent capabilities" 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] This is an early draft put together for fellowship applications, intended to give an overview of what I am currently working on. Results are presented for small sample sizes (n=30) and test-time compute budgets (up to 2048 tokens); the final manuscript will scale to the full datasets, and reasoning traces of approximately 16,000 tokens.Where appropriate, §blue text gives a flavour of expected findings once the study is complete.

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

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)?
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

2.1 Emergent capabilities of large language models

“Emergent capabilities" in language models has been used to refer to a range of related concepts, here, we use it to describe sudden, non-linear increases in model performance with respect to a scaling quantity (such as model parameters, training dataset size, or test-time budget). (Wei et al., 2022a; Srivastava et al., 2022) document a large number of model capabilities that display behaviour characteristic of “emergence"; for example (Wei et al., 2022a) show non-linear increases in performance on few-shot modular arithmetic, question and answer, and instruction following tasks when model training compute is scaled over multiple orders of magnitude. Following these findings, multiple studies sought to explain the reason for such behaviour. (Schaeffer et al., 2023) shows that when the scoring metric for model performance does not assign credit for partial progress towards a solution, sudden jumps in performance can result; however, changing the metric to reward correct steps towards a solution leads to much smoother scaling behaviour. For example, many capabilities identified as emergent in (Srivastava et al., 2022) use the Exact String Match metric, which only rewards a model when it produces a string identical to the target string. (Schaeffer et al., 2023) use Token Edit Distance instead, which scores the model based on the number of tokens in it’s output that must be altered to reach the target string. Other works posit a more natural view on emergent capabilities, arguing that highly compositional tasks are likely to exhibit sharp, non-linear scaling behaviour. (Arora and Goyal, 2023; Okawa et al., 2024; Yu et al., 2023). In short, when a model must be competent in several underlying skills before it can complete a task, it will score poorly on the task until the final skill is learnt (which could be done by scaling training examples, model size, training epochs, or another quantity), at which point it can suddenly become performant at the task. To illustrate, consider a model learning how to perform the task of 3-digit addition, and assume that the composite skills involved in this task are learning how to perform addition in the units, tens, and hundreds place-value columns. The model will need to learn all three skills to perform 3-digit addition, and upon learning how to add digits in the hundreds column performance on this task could increase rapidly. This phenomenon will be witnessed to a greater extent for tasks that have more complex compositional structure. (Okawa et al., 2024) illustrate this for visual concepts (such as object shape, size, and colour) generated by diffusion models, whereas (Lubana et al., 2024) introduce a formal language (specifically, a probabilistic context-sensitive grammar) as a toy model for the study of emergent capabilities in the natural language domain. The study tracks a model’s learning of underlying skills needed to generate sentences in this language, namely grammar acquisition, and the learning of type constraints (both relative and descriptive), and find that the points during training when the model learns and composes these skills leads to a sudden drops in loss of downstream tasks that is characteristic of emergent scaling behaviour. (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).

2.2 Test-time compute scaling language reasoning models

In recent years it has become increasingly popular to increase the test-time compute budget of a model to improve its performance. Arguably the simplest way to do so is sampling multiple responses from a model when prompted (Brown et al., 2025). Other methods include telling the model to produce a "chain of thought" (Wei et al., 2022b) prior to outputting it's final solution, as well as making the model self-verify it's reasoning as it progresses towards a final answer (Weng et al., 2023). These "reasoning" capabilities can be observed in standard LLMs, but the strongest models learn to reason through a combination of supervised fine-tuning on example reasoning traces (Muennighoff et al., 2025), and reinforcement learning (Guo et al., 2025). These language reasoning models (LRMs) display state-of-the-art performance on mathematics and logic benchmarks, and substantially outperform their "non-reasoning" counterparts (Epoch AI, 2024).

In response to the proliferation of LRMs, numerous works have looked to study the reasoning behaviour of LRMs more precisely, comparing model-generated reasoning traces to ones produced by humans (Marjanović et al., 2025), as well as the optimal allocation of compute to extend reasoning traces at the expense of producing a large number of responses in parallel (Snell et al., 2024). Notably, (Marjanović et al., 2025) finds that the reasoning traces ("thoughts") of DeepSeek-R1 follow consistent structure, performance follows a U-shaped curve when scaling reasoning trace length, and displays notable divergences with human-thinking processes.

3 Method

3.1 Test-time compute scaling method

Broadly there are two ways to scale test-time compute, sequential scaling and parallel scaling. Sequential scaling allows a model to produce a longer reasoning trace when responding to a prompt. Parallel scaling leverages additional compute to sample multiple outputs from a model, then aggregates these outputs using a method such as majority voting (Wang et al., 2022). 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[3].

To control the length of a reasoning trace, we follow (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[4], 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, we focus our studies on the following:

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 scaling behaviour on three medium-sized open-weight reasoning models:

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

[3]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.

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

The study aims to establish whether certain task distributions (benchmarks) and reasoning models are more likely to display emergent test-time scaling behaviour than others.

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.

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:

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

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 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:

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

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.

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.

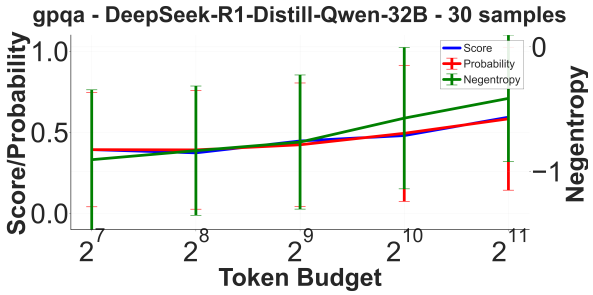

Figure 1: Scaling curves for score, probability of ground truth, and negative entropy over the solution set. Results presented for Deepseek-R1-Distill-32B on GPQA (30 samples).

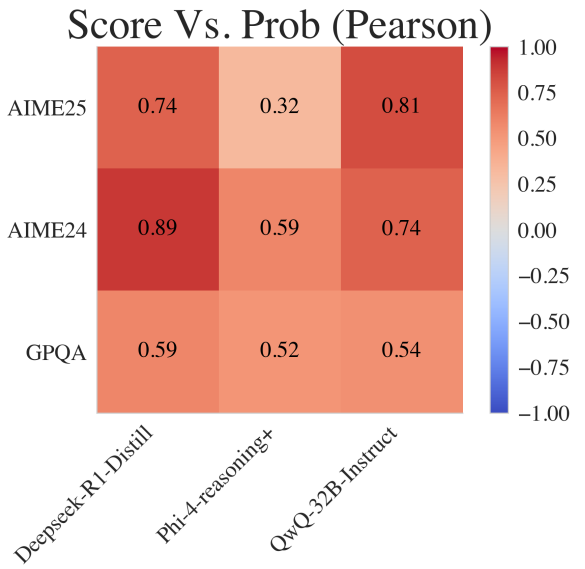

Figure 2: Pearson correlations between accuracy and ground truth probability (averaged over all samples) for all model-dataset pairings. As expected, moderate-to-strong correlations between discrete and continuous metrics are observed.

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.

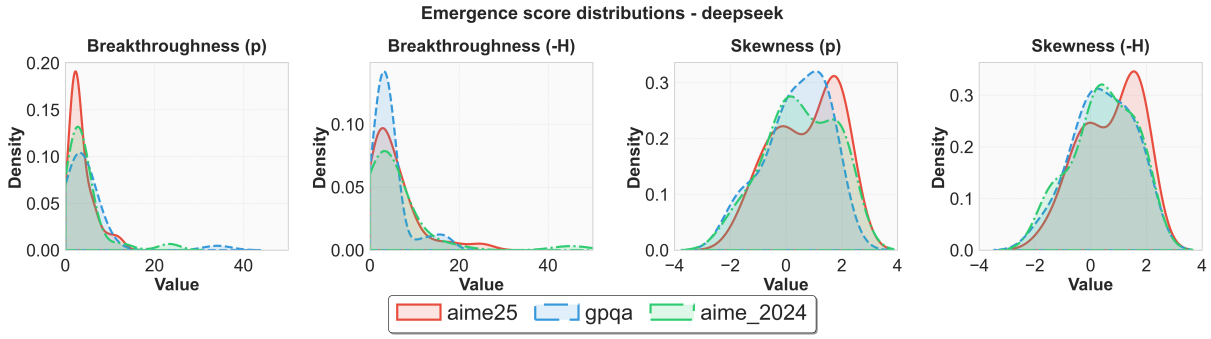

Figure 3: Full distributions of emergence scores across datasets when evaluated on Deepseek-R1-Distill-32B. Summary statistics are shown in Table 1. Distributions with modes shifted towards higher values indicate a greater degree of emergent test-time scaling.

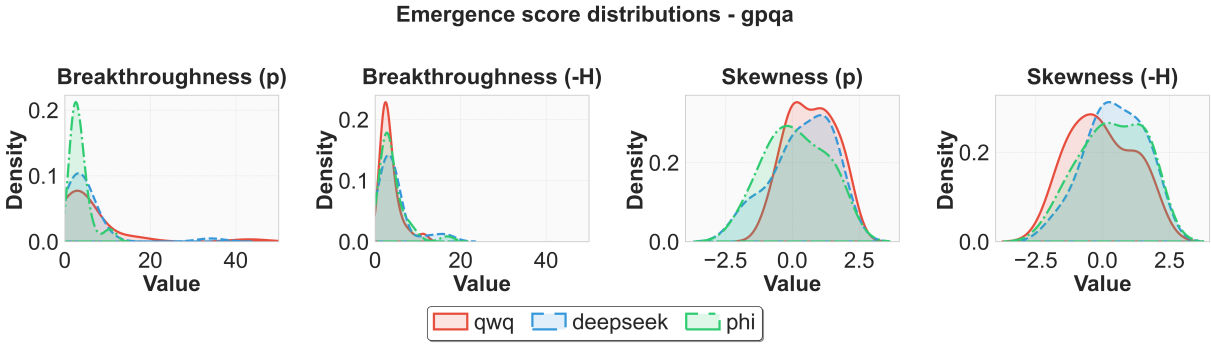

Figure 4: Full distributions of emergence scores across models when evaluated on GPQA. Summary statistics are shown in Table 2. Distributions with modes shifted towards higher values indicate a greater degree of emergent test-time scaling.

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.

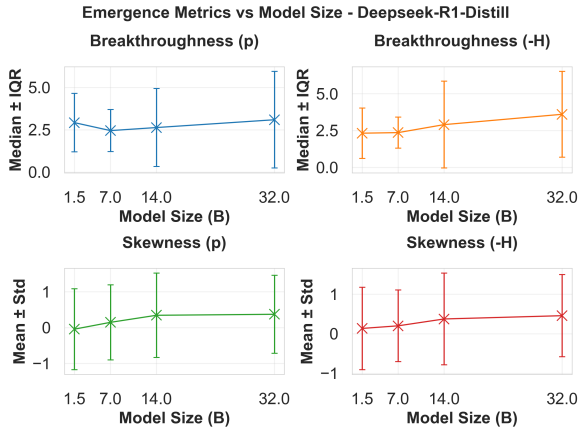

Figure 5: Distribution centre and spread statistics for the emergence scores (median and interquartile range for the breakthroughness metrics, mean and standard deviation for the skewness metrics) across Deepseek-R1-Distill model size. There is a general tendency for emergence scores to increase with 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.

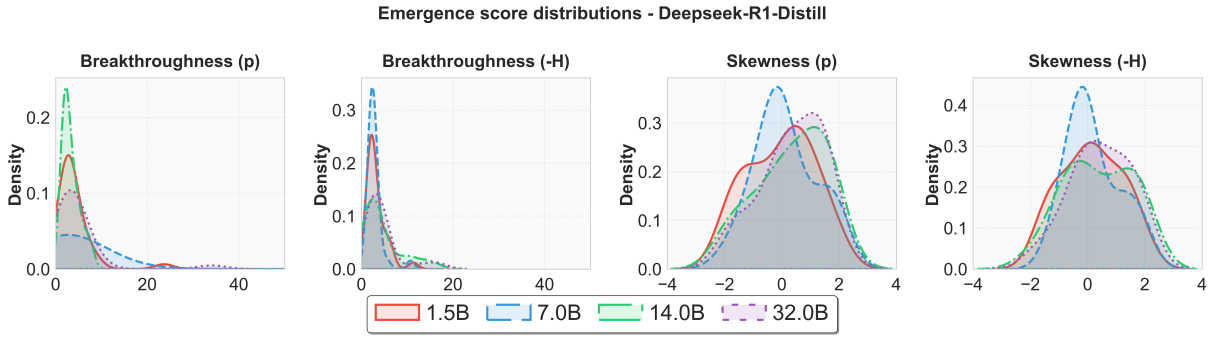

Figure 6: Distributions of emergence scores across Deepseek-R1-Distill model size. Summary statistics are shown in Figure 5. Distributions with modes shifted towards higher values indicate a greater degree of emergent test-time scaling.

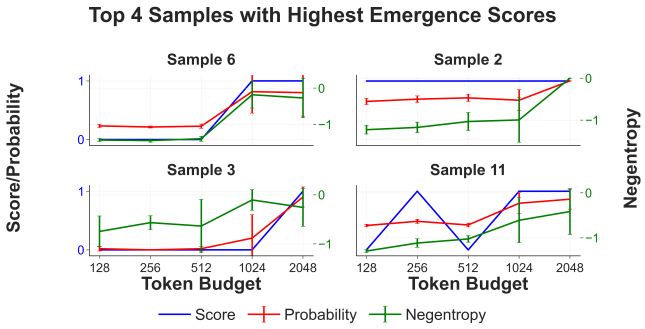

Figure 7: Samples from GPQA with highest emergence scores (when tracking the probability metric) when evaluated with Deepseek-R1-Distill-32B. Notice the non-linear scaling of ground truth probability with respect to token budget, which deviate from smooth, predictable scaling behaviour. This is especially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

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.

Epoch AI. 2024. "ai benchmarking hub". Accessed: 2025-09-24.

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.

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*.

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*.

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*.

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

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*.

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

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.

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*.

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, 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*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification.

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*.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. 2023. Skill-mix: a flexible and expandable family of evaluations for ai models.