Is Scaling Enough for Artificial General Intelligence?

Che-Wei, Hsu and Ming-Chun, Lee
Prof. Pei Wang
Topics in Computer Science
23 April 2026

Abstract

This report investigates whether scaling alone is sufficient to achieve Artificial General Intelligence (AGI). Scaling, which involves increasing model size, data, and computational resources, has driven major advances in modern AI systems such as GPT-4. By reviewing both supporting and opposing evidence from recent research, this report finds that while scaling can systematically improve performance and even lead to emergent capabilities, it does not fully address key challenges such as reasoning, truthfulness, and generalization. Studies like scaling laws and emergent abilities suggest promising progress, but findings from TruthfulQA and recent work on reasoning limitations reveal important weaknesses in large models. Therefore, this report concludes that although scaling is a powerful and necessary approach, it is not sufficient by itself to achieve true AGI, and additional innovations beyond scaling will likely be required.

1. Introduction

In recent years, Artificial Intelligence (AI) has developed very quickly, especially with large models like GPT-4. These models can perform many tasks, such as answering questions, writing code, and solving problems. A key reason for this progress is *scaling*, which means increasing the model size, the amount of data, and the computing power. In general, larger models tend to perform better on many tasks.

Because of these improvements, some researchers believe that scaling alone may eventually lead to Artificial General Intelligence (AGI). AGI refers to a system that can perform a wide range of tasks at a human level, not just one specific task. However,

even though current models are very powerful, it is still unclear whether they truly have general intelligence or just learn patterns from large amounts of data.

In this report, we will discuss the following question: **Is scaling enough to achieve AGI?** We will first introduce what scaling is, then discuss its advantages and limitations, and finally give my own opinion on whether scaling alone is sufficient.

2. What is Scaling?

Scaling, in the context of modern machine learning, refers to the systematic increase of three interdependent resources:

- **Model Size:** The number of trainable parameters. Larger models can learn more complex, higher-dimensional representations.
- **Training Data:** The volume and diversity of text (or other modalities) used during training. More data provides broader coverage of human knowledge and language.
- **Compute:** The total GPU/TPU processing power applied during training. Greater compute enables training larger models on more data within feasible time frames.

The foundational empirical insight is that model performance improves predictably as these three factors increase together, following a power-law relationship (Kaplan et al., 2020). This relationship implies that AI capabilities can be systematically engineered through scaling, offering a reliable and practical pathway for improving performance across a wide range of tasks.

3. Evidence Supporting the Scaling Hypothesis

A substantial body of research provides strong empirical support for the view that scaling may be a viable pathway toward Artificial General Intelligence (AGI).

3.1 Scaling Laws (Kaplan et al., 2020)

Kaplan et al. demonstrate that language model performance improves in a predictable power-law manner as model size, dataset size, and computational resources increase (Kaplan et al., 2020). These relationships hold across multiple orders of magnitude, indicating that performance gains can be systematically achieved through scaling. This predictability is particularly significant, as it enables researchers to estimate future performance prior to training. However, it is important to note that these results primarily concern improvements in language modeling loss rather than higher-level cognitive abilities such as reasoning or general intelligence. As a result, while scaling laws establish a strong empirical foundation, they do not by themselves demonstrate that scaling is sufficient for achieving AGI.

3.2 Emergent Abilities (Wei et al., 2022)

Wei et al. identify the phenomenon of emergent abilities, where certain capabilities appear only when models reach a sufficient scale (Wei et al., 2022). These include tasks such as multi-step reasoning and complex problem solving, where smaller models perform near chance while larger models achieve substantially higher accuracy. Unlike the smooth improvements predicted by scaling laws, these abilities represent qualitative shifts in model behavior. This suggests that scaling may not only improve performance quantitatively, but also enable the emergence of new capabilities. However, such abilities are often task-specific and do not yet constitute general intelligence.

3.3 Chinchilla: Principled Scaling (Hoffmann et al., 2022)

Hoffmann et al. refine the scaling hypothesis by demonstrating that many large models are undertrained relative to their size, and that optimal performance depends on balancing model parameters with training data (Hoffmann et al., 2022). Their results show that a smaller, properly trained model can outperform significantly larger models when trained on more data. This finding highlights that effective scaling is not simply a matter of increasing parameter count, but requires a principled allocation of computational resources. Consequently, scaling should be understood as an optimization problem rather than a purely quantitative expansion.

3.4 Sparks of AGI and GPT-4 (Bubeck et al., 2023)

Bubeck et al. provide evidence that large-scale models such as GPT-4 exhibit broad capabilities across domains including mathematics, coding, medicine, and law, often achieving performance close to human level (Bubeck et al., 2023). Notably, GPT-4 demonstrates the ability to generalize across diverse tasks without task-specific training, suggesting the emergence of flexible and transferable capabilities. These findings are further supported by reported benchmark results indicating near human-level performance on standardized evaluations. Together, these results suggest that sufficiently large and well-trained models can begin to approximate general-purpose problem-solving systems.

3.5 Summary of essay Supporting the Scaling Hypothesis

Taken together, these studies reveal a consistent pattern: scaling not only improves performance in a predictable and systematic manner, but can also lead to the emergence of new capabilities and more flexible behavior. This provides strong evidence that scaling is a promising pathway toward more general forms of intelligence. However, current evidence primarily demonstrates improvements in performance rather than the realization of full general intelligence, leaving open the question of whether scaling alone is sufficient to achieve AGI.

4. Evidence Against the Scaling Hypothesis

Despite the strong empirical support for scaling, a growing body of research highlights systematic limitations that scaling alone does not resolve.

4.1 Truthfulness Does Not Scale (Lin et al., 2021)

Lin et al. introduce the TruthfulQA benchmark to evaluate the truthfulness of language models and find that larger models do not necessarily produce more accurate answers (Lin et al., 2021). On this benchmark, the best-performing model achieves only about 58% truthfulness, compared to approximately 94% for humans. As summarized in the report, larger models tend to better imitate patterns in human-generated data, including misconceptions and falsehoods, rather than reliably producing correct

answers. This suggests that scaling improves fluency and plausibility, but does not guarantee factual reliability, which is a fundamental requirement for AGI.

4.2 The Illusion of Thinking (Apple, 2025)

Recent work by Apple investigates reasoning performance under controlled problem complexity and finds that model accuracy can collapse sharply as task difficulty increases (Apple, 2025). As noted in the analysis, models are often able to solve individual reasoning steps but fail to consistently combine them into coherent multi-step solutions when complexity grows. These findings indicate that current models rely heavily on pattern recognition rather than stable, generalizable reasoning processes, limiting their ability to handle complex or novel tasks.

4.3 Structural Limits: Informational Linkage Gaps (2026)

More recent theoretical work argues that scaling is fundamentally constrained by the structure of the training distribution (Montanino, 2026). The concept of informational linkage gaps refers to situations in which solving a problem requires connecting knowledge across domains in ways not explicitly represented in the data. As summarized in the report, such limitations are structural and cannot be overcome simply by increasing data or computational resources. This suggests that scaling primarily improves interpolation within known patterns, but does not enable the flexible abstraction required for general intelligence.

4.4 Measuring Intelligence vs. Skill (Chollet, 2019)

Chollet argues that current AI systems primarily demonstrate skill acquisition based on extensive exposure to data, rather than true general intelligence (Chollet, 2019). He defines intelligence as the ability to solve novel problems efficiently with minimal prior experience, emphasizing generalization and sample efficiency. Benchmarks such as the ARC (Abstraction and Reasoning Corpus) are designed to measure this capability, and large-scale models perform poorly on them relative to humans. From this perspective, scaling improves performance on familiar tasks but does not necessarily lead to genuine general intelligence.

5. The Central Debate

The debate over scaling ultimately centers on a fundamental question: does scaling lead to qualitative general intelligence, or does it merely produce quantitative improvements in pattern recognition?

Supporters of the scaling hypothesis argue that continuous performance gains across a wide range of benchmarks, combined with the emergence of new capabilities at larger scales, suggest that general intelligence may arise as a natural consequence of sufficient scale. From this perspective, intelligence may emerge gradually as models become more powerful, and current limitations primarily reflect incomplete engineering. Models may not yet be large enough or trained on sufficiently diverse data, rather than being constrained by fundamental limitations of the underlying approach.

In contrast, critics argue that existing evidence points to persistent and structural limitations. While large models achieve strong performance on aggregate benchmarks, they continue to struggle with key aspects of intelligence, including truthfulness, robust reasoning, and generalization to novel problems. Improved benchmark performance does not necessarily indicate genuine intelligence, but may instead reflect increasingly advanced forms of pattern recognition rather than true understanding.

Therefore, the central question is whether scaling can produce a qualitative shift toward general intelligence, or whether it will ultimately plateau as an increasingly effective but fundamentally limited form of statistical pattern recognition.

6. Our Position

Based on the analysis above, this report argues that scaling alone is not sufficient to achieve Artificial General Intelligence (AGI). While scaling has significantly improved model performance and enabled systems to handle a wide range of tasks, these improvements do not necessarily indicate the emergence of true general intelligence.

Empirical evidence suggests that large-scale models can demonstrate impressive cross-domain capabilities, as shown in studies of GPT-4 and related systems (Bubeck et al., 2023). However, these capabilities do not imply genuine understanding. At the same time, multiple lines of research reveal persistent limitations. For example, TruthfulQA shows that larger models may produce less truthful outputs despite

improved overall performance (Lin et al., 2021). Similarly, recent work on reasoning demonstrates that model performance degrades as task complexity increases, indicating a lack of stable and generalizable reasoning processes (Apple, 2025).

Taken together, these findings suggest that scaling primarily enhances pattern recognition rather than addressing deeper challenges such as reasoning, abstraction, and generalization. Since AGI requires the ability to adapt to novel situations and perform flexible, compositional reasoning, these limitations indicate that scaling alone is unlikely to be sufficient. Achieving AGI will likely require new paradigms or architectural innovations beyond simply increasing model size, data, and compute.

7. Conclusion

In conclusion, scaling has played a central role in the rapid advancement of artificial intelligence, driving consistent improvements in performance across a wide range of tasks. Research on scaling laws and emergent abilities demonstrates that increasing model size, data, and computational resources can significantly enhance model capabilities.

However, these improvements do not necessarily translate into general intelligence. As discussed, large models continue to exhibit limitations in truthfulness, reasoning, and generalization, suggesting that scaling primarily improves performance within the scope of existing data distributions rather than enabling fundamentally new forms of intelligence.

Overall, scaling should be understood as a powerful but incomplete approach. While it provides a strong foundation for progress, it does not fully address the core requirements of AGI. Future research will likely need to move beyond scaling and explore new methods that can better support reasoning, abstraction, and adaptive learning.

References

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4.* arXiv:2303.12712.

Chollet, F. (2019). *On the Measure of Intelligence.* arXiv preprint arXiv:1911.01547.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. (2022). *Training compute-optimal large language models.* arXiv:2203.15556.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). *Scaling laws for neural language models.* arXiv:2001.08361.

Lin, S., Hilton, J., & Evans, O. (2021). *TruthfulQA: Measuring how models mimic human falsehoods.* arXiv:2109.07958.

OpenAI. (2023). *GPT-4 Technical Report.* arXiv:2303.08774.

Shojaee, M., Mirzadeh, I., Alizadeh, K., Horton, M., Farajtabar, M., & Bengio, S. (2025). *The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity.* Apple Machine Learning Research. arXiv:2506.06941.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). *Emergent abilities of large language models.* Transactions on Machine Learning Research. arXiv:2206.07682.