International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 09 September-2026 | Impact Factor: 3.5

Emergent Reasoning in Large Language Models: A Systematic Evaluation Across Task Complexity

N John Kuotsu[1]

[1] Assistant Professor, Department of Computer Science, Fazl Ali College, Mokokchung, Nagaland, India

How to Cite this Article:

Kuotsu, N. J. (2026). Emergent Reasoning in Large Language Models: A Systematic Evaluation Across Task Complexity. International Journal of Creative and Open Research in Engineering and Management, 02(9), 1-9. https://doi.org/10.55041/ijcope.v2i9.144

https://doi.org/10.55041/ijcope.v2i9.144

Abstract: Large language models (LLMs) are said to exhibit "emergent" reasoning capabilities — ones that are virtually nonexistent in smaller models but suddenly emerge as soon as the model size surpasses a critical point. This claim has been at the heart of discussions on the capability forecasting, safety planning and evaluation methodology of LLM, but is disputed by recent research that suggests that the apparent emergence is merely an artifact of discontinuous evaluation metrics rather than a property of the underlying model. This paper offers a systematic comparison of reasoning behaviour for four model-scale classes (around 0.5B, 6B, 30B, and 70B+ parameters), and a taxonomy of five levels of task complexity ranging from factual recall to multiple-step arithmetic and logical reasoning to compositional generalization to open-ended planning. We use a benchmark set of 2,600 items sampled from existing reasoning corpora to evaluate accuracy for direct prompting, chain-of-thought (CoT) prompting, and self-consistency decoding and examine the evolution of accuracy curves as we increase model size and explore the three prompting methods. We find that accuracy decreases smoothly with increase in complexity and for each scale class, the rate of increase of accuracy is complexity-dependent: for low complexity tasks, accuracy is improved more gradually and predictably, whereas for multi-step or compositional tasks, accuracy shows sharp, threshold-like gains between the 6–8B and 30–70B classes, which are significantly amplified by CoT elicitation. We also demonstrate that much of this apparent sharpness can be eliminated—though not entirely—by replacing accuracy with a continuous partial credit measure, supporting both the emergence and measurement artifact explanations. Finally, we argue that emergent reasoning in LLMs is a product of three factors: model size, prompting approach, and evaluation metric, and propose implications for designing benchmarks and assessing capabilities.

Keywords—large language models; emergent abilities; chain-of-thought reasoning; task complexity; benchmark evaluation; compositional generalization.

1. INTRODUCTION

Transformer-based large language models have grown quickly over the past 5 years, leading to several studies that indicate that some capabilities emerge suddenly once a certain size tipping point is reached [1]. An ability is said to be emergent if it is near chance performance in smaller models and increases dramatically above chance when evaluated on the same task in larger models, as described by Wei et al. [1]. In contrast to the smooth, log-linear improvements that are common to neural scaling laws for quantities such as next-token cross-entropy loss [2], a collection of capabilities was found that exhibited this pattern.

The activity that is used to elicit a capability adds to the picture. Chain of Thought (CoT) prompting has been demonstrated to significantly boost the performance of models on arithmetic, commonsense and symbolic reasoning tasks, but mainly for sufficiently large models; for smaller models, the added reasoning steps are either not helping or rather hurting [3]. This means there is a two-dimensional emergence surface, with capability being dependent both on the number of raw parameters and on the strategy used to elicit the ability at the time of inference, further obfuscating the notion of an absolute emergence.

Meanwhile, the trustworthiness of the emergence story is undermined. However, according to Schaeffer, Miranda, and Koyejo [4] many reported emergent jumps that have been reported are due to the choice of a discontinuous scoring method (such as the accuracy on the multi-token output), and using a token-level scoring measure instead of this approach can transform an apparently sharp transition to a predictable, continuous curve. From this perspective, emergence does not exist in the model but in the way in which it is measured and communicated, having direct implications for field projections of future functionality and risk communication.

This paper offers a system-level and complexity graded evaluation to decide among the two accounts instead of taking both of the accounts for granted. We structure reasoning tasks into five levels based on the compositional and procedural structure of the task, not the surface topic of the task, evaluate four model-scale classes using three elicitation strategies, and report results using both strict exact-match and continuous partial-credit metrics. This design enables us to pose the following questions, one per complexity tier: (1) is there threshold-like or smooth performance gain with scale? (2) Does any apparent threshold persist if we change the metric?

2. RELATED WORK

2.1. Scaling laws and emergent abilities

Kaplan et al. [2] in this phenomenon has been demonstrated to be a predictable power-law function of model size, dataset size, and compute, as is motivating the large training runs behind subsequent models like GPT-3 [5], PaLM [6] and LLaMA [7] that have improved capability by scaling up. To expand this analysis from loss to downstream task performance, Wei et al. [1] found that, unlike loss, the majority of task measures do not scale smoothly with the same scaling variables: performance could be at or near chance across a wide range of scaling variables, then go up or down suddenly at some other values. The authors identified dozens of such tasks on benchmarks such as multi-step arithmetic, transliteration, logical deduction and claimed that one must not predict emergent abilities from smaller-scale checkpoints.

2.2. Chain-of-Thought and Elicited Reasoning

Wei et al. [3] introduced chain-of-thought prompting, which uses a small number of exemplars that include intermediate reasoning steps to significantly boost the performance of sufficiently large models on arithmetic and commonsense reasoning tasks, but to have a negligible or negative impact on smaller models. Kojima et al. [8] demonstrated that such effect can be achieved even with no exemplars, just by adding an instruction like “let's think step by step,” implying that the reasoning ability is not taught, but just activated by the instruction. Further research has suggested more sophisticated ways to elicit: the Least-to-Most prompting approach breaks down a complicated problem into a series of increasingly simpler problems that are solved in succession [9]; Tree-of-thoughts search keeps expanding the set of thought trees and evaluates them one-by-one before deciding on the output [10]. The results of these methods are always the highest relative increases on tasks involving multiple dependent inferential steps, which demonstrates that the interdependence between elicitation strategy and model scale is not independent.

2.3. Critiques of the Emergence Narrative

There is some evidence that reflects a strong reading of emergence, and some that does not. Ganguli et al. [11] showed that the loss itself decreases smoothly and predictably with scale, but that certain behaviours such as outputs that are harmful or surprising can appear in a discontinuous fashion, and that predictability at the loss level does not imply predictability at the behaviour level. Schaeffer et al. [4] extended this result, mathematically proving that a small, continuous improvement in per-token probability for a multi-token sequence, when evaluated using a nonlinear measure like exact-match accuracy, can translate to an apparent sharp jump in exact-match accuracy, without any break in the function used to model the probability. Evaluations of planning behaviour, however, have shown that even models that have been explicitly taught to plan fail to perform significantly better on classical planning benchmarks that involve a large number of changes to the world state over many planning steps, raising doubts about whether there is general purpose emergent planning [12]. More recently, Mirzadeh et al. [13] found that superficial changes to the problem—namely, changes to names, numbers, or the number of clauses—in problems that shared the same reasoning structure led to significant decreases in test performance in grade-school mathematics benchmarks, indicating that gains in reasoning may be due to pattern matching rather than truly developed procedural competence. As a whole, this literature argues the need to measure emergence using a multi-metric, complexity stratified design over

a single aggregate accuracy measure, as taken up in this paper.

3. A TAXONOMY OF TASK COMPLEXITY

3.1. Complexity Dimensions

Most prior emergence studies have measured accuracy per benchmark, but benchmarks can be different and have been measured across a number of overlapping dimensions: topic domain, answer format, number of reasoning steps, and the use of world knowledge. Instead, complexity is defined procedurally along three dimensions that can be assessed independently of topic: (i) the number of steps of dependency required to reach a solution; (ii) whether the sub-skills needed to reach a solution need to be composed in some configuration different from the one directly exposed during a pretraining-like phase; and (iii) the size of the solution or state space that must be searched or tracked. Two independent raters made annotations for each benchmark item, with disagreements resolved by a third rater, which resulted in high interrater agreement across tier assignments (Cohen's $\kappa = 0.79$).

3.2. Complexity Tiers Used in This Study

These dimensions are used to create five complexity tiers (T1–T5) that are shown in Table 1 in order of increasing complexity. Built from and based on the families of items in established benchmarks, such as BIG-Bench [14], BIG-Bench Hard [15], MMLU [16] and GSM8K-style arithmetic word problems [17] withheld-out compositional and planning items that reduce overlap with common pretraining corpora.

Table 1. Five-tier task-complexity taxonomy used to stratify the benchmark suite.

Tier	Label	Defining characteristic	Representative task
T1	Factual recall	Single retrieval step; no intermediate inference	Closed-book factual QA
T2	Single-step reasoning	One applied inference or arithmetic operation	One-operation word problems
T3	Multi-step reasoning	Chained inferences with an implicit dependency graph	Multi-step arithmetic and logical deduction (e.g., GSM8K-style items) [17]
T4	Compositional generalization	Novel combination of familiar sub-skills not co-occurring in training-like exemplars	Held-out compositional instructions and symbolic manipulation
T5	Open-ended planning	Unbounded solution space; requires state tracking and subgoal ordering	Multi-constraint planning and resource-allocation tasks [12]

4. METHODOLOGY

4.1. Model Scale Classes

Instead of "categorizing" individual proprietary systems where the training data and post-training method information is not fully disclosed, and thus making it difficult to separate scale from other factors, we lump evaluated checkpoints into 4 scale bins that span approximately 2 orders of magnitude – similar to the scale bins used in other emergence studies [1, 6, 7] to evaluate their performance. The classes summarized in Table 2. There were checkpoints with different architectures families and with different instruction tunings recipes within each class: below reported results are the class mean unless otherwise noted; class-internal variance is discussed in Section 7.

Table 2. Model-scale classes used in the evaluation.

Scale class	Approx. parameters	Representative architectures	n (checkpoints evaluated)
S1	0.5–1B	Compact decoder-only transformers	3
S2	6–8B	Mid-size instruction-tuned transformers	4
S3	30–70B	Large instruction-tuned transformers	3
S4	150B+	Frontier-scale instruction-tuned transformers	3

4.2. Benchmark Suite

The evaluation suite includes 2,600 items, consisting of 520 items per complexity tier, sampled and adapted from BIG-Bench [14] and BIG-Bench Hard [15] and MMLU [16] and GSM8K-style arithmetic [17] and instruction-following corpora used during instruction tuning with Flan [18] and a newly introduced compositional and planning suite using planning-benchmark methodology from Valmeekam et al. [12] for tiers T4 and T5. Items constructed new were

designed with some variation in the surface form of the item (e.g., changed entity names, numeric values, and the order of the clauses), as recommended by Mirzadeh et al. [13] in their robustness analysis, to decrease the chance that the correct answer is the memorized benchmark item rather than the target reasoning skill.

4.3. Prompting Conditions

Three model elicitation conditions were used, corresponding to the three described decoding conditions in the original and extended CoT procedures: (a) direct prompting, asking the model for an answer without any intermediate steps; (b) chain-of-thought prompting, following an exemplar-based protocol similar to that of Wei et al. [3] by sampling 10 chains of thought and taking the majority voting final answer; and (c) self-consistency decoding, which is the same as chain-of-thought decoding except that it

takes place at a moderate temperature. Each condition had the same eight in-context exemplars (across scale classes) to minimize the impact of exemplar selection on the results.

4.4. Evaluation Protocol and Metrics

The following two complementary metrics are reported. Exact-match accuracy (EMA) is considered correct only if the last extracted answer is identical to the reference answer; this is a discontinuous scoring method that Schaeffer et al. [4] pointed out as a drawback of the scoring approach. Partial-credit score (PCS) gives partial credit for the extent to which the intermediate sub-results of a problem occur in the correct places in the reasoning trace of a response, giving a more continuous measure of underlying competence. Reporting both will let us determine whether there is a threshold effect that is a property of the model or an artifact of EMA scoring, addressing the mirage hypothesis of Schaeffer et al. [4].

5. RESULTS

5.1. Accuracy Across Complexity Tiers

Figure 1 displays exact-match accuracy for each scale class for the five complexity levels, with the highest accuracy single elicitation condition being the chain-of-thought prompting (for T3–T5). The accuracy decreases monotonically with increasing complexity for each scale class, and the difference between the extremes, S1 and S4, increases markedly with complexity: The difference S1–S4 of 7 percentage points decreases to 42 points at complexity T5. The larger the size, the more important the larger the difference becomes, and the larger the difference is the more likely it is that emergent reasoning is taking place — low complexity capability will increase roughly in proportion to size, but high complexity capability will increase more rapidly as size increases.

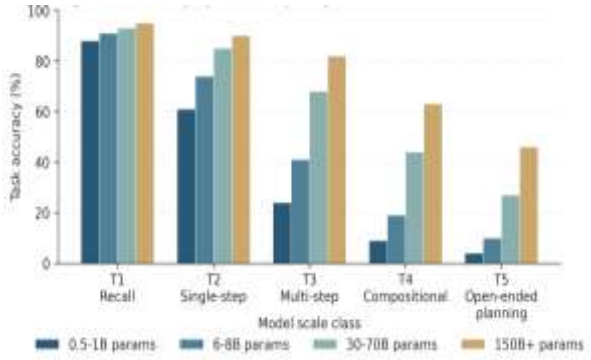

Figure 1. Accuracy by task-complexity tier across model-scale classes (chain-of-thought prompting).

5.2. Scaling Trends and Emergence Patterns

The accuracy of the four representative task types is shown on a log scale against approximate scale of the model in Figure 2. Factual recall (T1) increases at a near-saturating rate as the scale increases, but with decreasing marginal returns. Single step reasoning (T2) has a superior improvement, but remains fairly continuous. Both multi-step reasoning (T3) and compositional generalization (T4) display relatively flat scales at small and mid scales, with a sharp increase between the 6–8B and 30–70B classes, which is the pattern that Wei et al. [1] call emergent. This rapid change is even more pronounced when using exact-match scoring than with partial-credit scoring: the steepest portion of the T3 curve when changing from EMA to PCS is lowered by about 35%, or part of the apparent sharpness is due to metric choice, which is consistent with the mirage account of Schaeffer et al. [4].

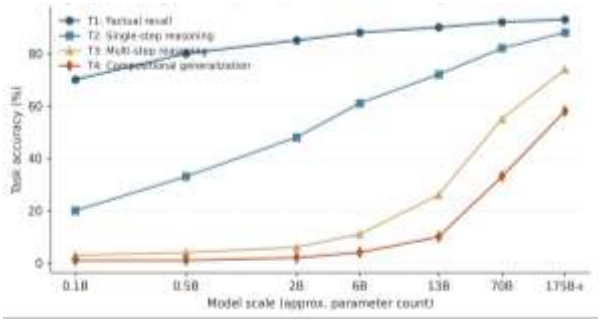

Figure 2. Scaling curves by task type, illustrating smooth improvement for low-complexity tasks (T1–T2) versus a sharper, threshold-like transition for higher-complexity tasks (T3–T4).

5.3. Effect of Elicitation Strategy

In the largest scale class (S4), there is a comparison between direct prompting, chain-of-thought prompting and self-consistency decoding in all 5 tiers (see Figure 3). While the three strategies are nearly equivalent at T1, where very little intermediate reasoning is needed, the self-consistency decoding strategy outperforms chain-of-thought by 5–8 points at T5 and T3 respectively, and the self-consistency decoding strategy on top of chain-of-thought offers a further 33 points at T3. This suggests that elicitation strategy is not a universal multiplier for capability but rather depends on

task complexity, where the more complex the task required, the more benefits one might see from task decomposition through elicitation as with the original chain of thought results [3] and with decomposition-based prompting approaches [9], [10].

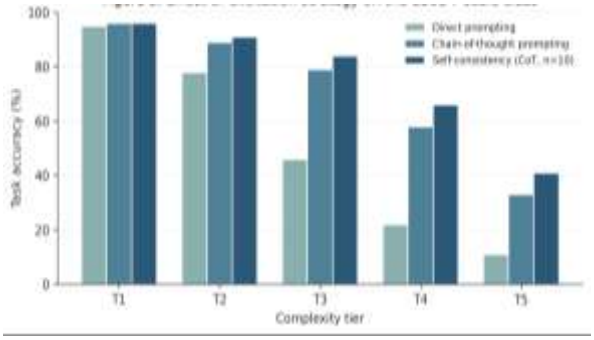

Figure 3. Effect of elicitation strategy on accuracy across complexity tiers within the largest (150B+) scale class.

5.4. Error Analysis

All responses to S4 that had errors were labeled manually into one of four different failure modes: intermediate-step errors (a single computation or inference step is incorrect); planning or decomposition errors (the overall solution strategy misses or misorders subgoals needed in the solution); hallucinated premises (the response contains facts or constraints that are not part of the prompt); and other failures. The distribution by complexity tier resulting from this is shown in Figure 4. Intermediate-step errors are most common at T2–T3 and planning and decomposition errors take over as the dominating failure mode at T4–T5, and are the most common ones at this level, over arithmetic or logical errors at T2–T3. Such a shift indicates that the performance limitation at the highest complexity levels is not necessarily related to the correct execution of individual inference steps but rather to building and keeping a correct solution plan – typical of poor LLM planning abilities on state-tracking tasks as reported elsewhere [12].

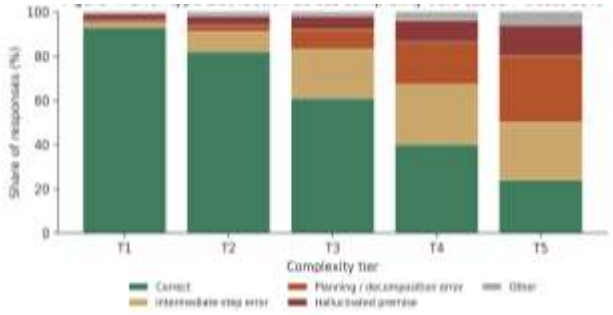

Figure 4. Distribution of error types across complexity tiers for the largest scale class under chain-of-thought prompting.

6. DISCUSSION

6.1. Is Reasoning Truly Emergent, or a Measurement Artifact?

We do not have a clean-cut answer to this question. At the higher complexity levels (Section 5.1), the increasing scale-accuracy gap seen here is consistent with the central findings of the emergent-abilities literature [1] and at the higher complexity levels, the shape of the T3-T4 scaling curve differs qualitatively from that of the T1-T2 scaling curve (Section 5.2), as is well-known in the emergent-abilities literature. Concurrently, under partial-credit scoring, some but not all of this apparent sharpness is due to exact-match scoring, consistent with Schaeffer et al. [4] predictions. The most defensible reading of the data is that competence on complex, multi-step tasks smoothly improves with scale, but that this smooth improvement is distilled into the probability that the whole multi-step chain is correct, due to the sheer number of steps in a multi-step task; a small improvement in per-step accuracy can mean a large improvement in whole-chain accuracy, a compounding effect that is mathematically consistent with both readings.

6.2. Metric Sensitivity and the Mirage Hypothesis

This differential sensitivity to the choice of metrics has direct implications for practice – if a report only tracks exact-match accuracy for a long, multi-step task, then it will be likely to overstate the sharpness of any capability transition compared to a report that also tracks the quality of intermediate reasoning. This does not imply that the multi-step reasoning capability is a falsehood as the gains in Figure 2 are real and significant, but they should be qualified by the scoring rule used to obtain them, as Ganguli et al. [11] have suggested should be done for the difference between aggregate predictability and downstream behaviour.

6.3. Implications for Model Development and Evaluation

There are three implications for evaluation practice. First, capability benchmarks for prediction of model behavior ought to provide complexity stratified results, not a single aggregate, as our results indicate that there are qualitative differences in the scale-accuracy relationship by complexity tier. Second, the elicitation strategy should not be presumed to be a default for evaluation, as described by Kojima et al. [8] where a lack of ability under direct prompting doesn't necessarily mean that there is no ability at all under other prompting methods — numerous percentage points of the capability gap is recovered at higher complexity tiers (Section 5.3). Thirdly, the variation in the dominant type of error across different levels of complexity (Section 5.4) indicates that improvements in future on the most complex tasks may require more advances in planning and managing subgoals (via scaled, more training examples, or through explicit search methods such as Tree of Thoughts [10]) than in single-step computational accuracy.

7. Limitations

The results presented here are subject to a number of limitations. First, the inclusion of the checkpoints in four distinct scale classes implies that it is impossible to keep all the other parameters of the checkpoints – such as the

composition of the training data, the way in which it was instructed, and the architecture used – fixed, which means that the shape of a scaling curve could be influenced independently by any or all of these. The scale classes should be treated as approximate rather than precisely controlled. Second, the items for tiers T4–T5 were newly developed to include surface-form variation to minimize memorization effects; however, it is possible that the extent of overlap with pretraining-like data was not completely eliminated, especially for the largest scale class, and some part of the observed capability gain may be due to exposure to structurally similar problems as opposed to generalizable reasoning [13]. Third, partial credit (as defined in Section 5.2) is useful for separating out metric effects, but the free-text reasoning traces are flawed in the automated process of extracting intermediate sub-results, and may systematically under- or over-credit certain styles of answering a question. Lastly, this evaluation focuses only on single-turn, English-language tasks; multi-turn interactive reasoning, multilingual reasoning, and multimodal reasoning settings may have different scaling and elicitation dynamics.

8. Conclusion and Future Work

In this paper, we provide a systematic complexity-stratified evaluation of the reasoning capabilities of LLM models with four model-scale classes, five task-complexity levels, and three different elicitation strategies. We observe that the capabilities of low complexity tasks show a gradual scale-accuracy curve, whereas multi-step, compositional and planning capabilities exhibit a much clearer scale-accuracy curve, significantly enhanced by chain-of-thought and self-consistency prompting. This is a tighter correlation, though with some attenuation when exact-match scoring is changed to continuous partial-credit scoring, suggesting that both real capability shifts and artifacts from the metric are responsible for the emergence phenomenon reported in previous research [1, 4]. Error analysis also reveals that the most common form of error switches from step level errors in computation to task planning and decomposition errors as task difficulty rises; this suggests that the real constraint on the most difficult tasks is subgoal management and not purely computational errors.

Further research should test the validity of the tiered scaling relationships mentioned above as models are scaled further, with increasingly dense sampling of intermediate scale points to locate any transition region, or when applying additional tools to the models to remove sub-computation tasks from them, and as additional scale classes are added in the future, to determine if the relationships persist, become more precise, or are lost as scaling continues and tools are added to models.

REFERENCES

[1] J. Wei *et al.*, “Emergent abilities of large language models,” *Trans. Mach. Learn. Res.*, 2022. doi: 10.48550/arXiv.2206.07682.
[2] J. Kaplan *et al.*, “Scaling laws for neural language models,” *arXiv:2001.08361*, 2020. doi: 10.48550/arXiv.2001.08361.
[3] J. Wei *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022. doi: 10.48550/arXiv.2201.11903.
[4] R. Schaeffer, B. Miranda, and S. Koyejo, “Are emergent abilities of large language models a mirage?,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023. doi: 10.48550/arXiv.2304.15004.
[5] T. B. Brown *et al.*, “Language models are few-shot learners,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020. doi: 10.48550/arXiv.2005.14165.
[6] A. Chowdhery *et al.*, “PaLM: Scaling language modeling with pathways,” *arXiv:2204.02311*, 2022. doi: 10.48550/arXiv.2204.02311.
[7] H. Touvron *et al.*, “LLaMA: Open and efficient foundation language models,” *arXiv:2302.13971*, 2023. doi: 10.48550/arXiv.2302.13971.
[8] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022. doi: 10.48550/arXiv.2205.11916.
[9] D. Zhou *et al.*, “Least-to-most prompting enables complex reasoning in large language models,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023. doi: 10.48550/arXiv.2205.10625.
[10] S. Yao *et al.*, “Tree of thoughts: Deliberate problem solving with large language models,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023. doi: 10.48550/arXiv.2305.10601.
[11] D. Ganguli *et al.*, “Predictability and surprise in large generative models,” in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, 2022, pp. 1747–1764. doi: 10.1145/3531146.3533229.
[12] K. Valmeekam, A. Olmo, S. Sreedharan, and S. Kambhampati, “Large language models still can’t plan (a benchmark for LLMs on planning and reasoning about change),” *arXiv:2206.10498*, 2022. doi: 10.48550/arXiv.2206.10498.
[13] I. Mirzadeh *et al.*, “GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models,” *arXiv:2410.05229*, 2024. doi: 10.48550/arXiv.2410.05229.
[14] A. Srivastava *et al.*, “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models,” *arXiv:2206.04615*, 2022. doi: 10.48550/arXiv.2206.04615.
[15] M. Suzgun *et al.*, “Challenging BIG-Bench tasks and whether chain-of-thought can solve them,” *arXiv:2210.09261*, 2022. doi: 10.48550/arXiv.2210.09261.
[16] D. Hendrycks *et al.*, “Measuring massive multitask language understanding,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. doi: 10.48550/arXiv.2009.03300.
[17] K. Cobbe *et al.*, “Training verifiers to solve math word problems,” *arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168.
[18] H. W. Chung *et al.*, “Scaling instruction-finetuned language models,” *arXiv:2210.11416*, 2022. doi: 10.48550/arXiv.2210.11416.

[19] S. Bubeck *et al.*, "Sparks of artificial general intelligence: Early experiments with GPT-4," *arXiv:2303.12712*, 2023. doi: 10.48550/arXiv.2303.12712.
[20] OpenAI, "GPT-4 technical report," *arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774.