Are Emergent Abilities of Large Language Models a Mirage?

Paper Summary

Chip Henderson - 48996654

SMU CS8321, 24 January 2024

The selection of appropriate metrics is one of the most important components of performance assessment. Schaeffer, Miranda, and Koyejo, the authors of the paper, *Are Emergent Abilities of Large Language Models a Mirage?*[1] go to great lengths to prove their hypothesis correct: that Large Language Models do not, in fact, suddenly manifest emergent abilities simply because of their size. Further, they postulate the arrival at such a conclusion is the result of an insufficient quantity and diversity of metrics.

The author's quote the definition of emergent abilities in the paper *Emergent Abilities of Large Language Models*, as "abilities that are not present in smaller-scale models but are present in large-scale models; thus they cannot be predicted by simply extrapolating the performance improvements on smaller scale models" [2]. These properties fall into categories of sharpness and unpredictability in model outputs. Schaeffer, Miranda, and Koyejo set out to disprove this assessment with a series of tests described below:

> We present our alternative explanation in a simple mathematical model, then test it in three complementary ways: we (1) make, test and confirm three predictions on the effect of metric choice using the InstructGPT/GPT-3 family on tasks with claimed emergent abilities; (2) make, test and confirm two predictions about metric choices in a meta-analysis of emergent abilities on the Beyond the Imitation Game Benchmark (BIG-Bench); and (3) show how to choose metrics to produce never-before-seen seemingly emergent abilities in multiple vision tasks across diverse deep network architectures. [1]

To acquire an understanding of the approaches used for testing their hypotheses, additional research was required. Each method proposed applied various mechanisms to compare model performance. Item one required the application of different metrics, continuous and linear, to models similar to those in the original claim. Item two utilized a benchmarking tool called "BIG-bench," a GitHub repository of language model benchmarks, introduced in the paper *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*[3] to confirm "predictions on the effect of metric choice"[1].

Item three required performing similar tests in an unrelated model architecture to prove emergent abilities could be artificially manifested in models simply through metric selection.

The first approach strives to show how "smooth, continuous, predictable changes in model family performance appear sharp and unpredictable"[1] simply through metric selection. Supporting this argument, the authors focus on the selection linear vs non-linear, or discontinuous metrics. Specifically, non-linear metrics such as accuracy in Machine Learning Models may show significant differences from linear metrics due to the "neural scaling laws: empirical observations that deep networks exhibit power law scaling in the test loss as a function of training dataset size, number of parameters or compute"[schaeffer2023are]. The original stipulation that emergent capabilities appear in larger models proposed by Wei, et. al. used accuracy as a metric[2]. Accuracy can be considered non-linear because small changes in the model or system can lead to non-linear changes in results. The input is not always equal to the output as a linear metric would show[4]. When similar models were tested using continuous metrics such as Accuracy and Multiple Choice Grade, the same sharp increases indicating emergent capabilities were noted. However, when linear metrics such as Token Edit Distance were employed the results indicated no emergent capabilities[1].

The second approach showed that when using a wider selection of metrics some aspects of Large Language Models show emergence, while a far larger number do not. To do this, they employed a benchmark analysis for Large Language Models known as BIG-Bench against the GPT family of models. BIG-Bench is a "large-scale, extremely difficult and diverse benchmark". It consists of 204 or more language tasks that challenge a model based on parameters such as correctness, thoroughness, novelty, and problems being "not solvable by memorizing the internet"[3]. Consistent use of benchmarks is an important construct to evaluate performance because it allows like-kind comparison of seemingly dissimilar systems. When the authors applied the BIG-Bench benchmark against the GPT family of models they found "emergent abilities appear only for specific metrics, not task-model families". Further, the authors analyzed "hand-annotated

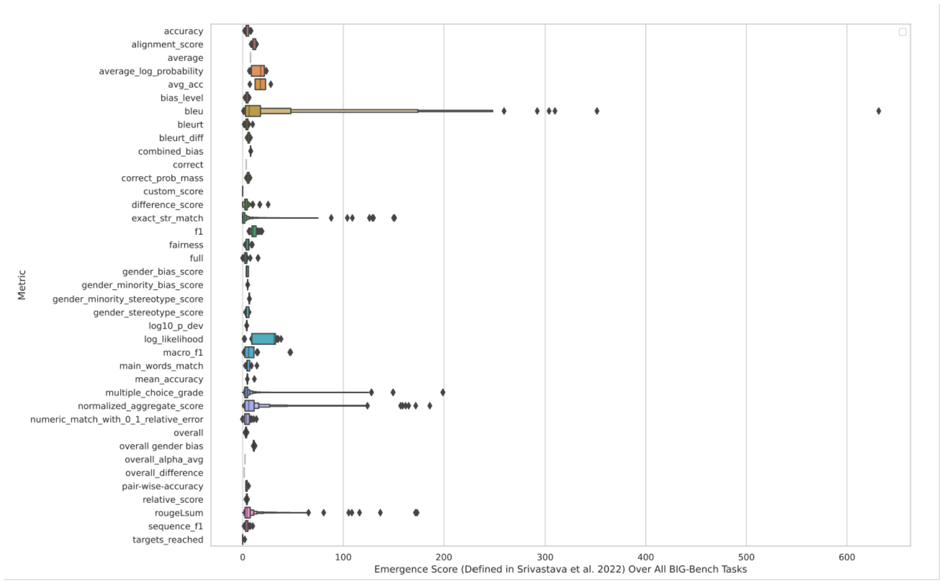

Figure 1: BIG-Bench Results on GPT Family of Models[1]

task-metric-model family triplets, which revealed emergent abilities appear with 4/39 metrics, and 2 metrics account for $> 92\%$ of claimed emergent abilities"[1].

The third approach to assess the credibility of the claim involved the use of multiple vision tasks and deep network architectures to "force" emergent capabilities to appear. The choice of vision tasks by the authors was to prove they could intentionally produce the appearance of emergent capabilities based on metric selection in an area that had not previously shown that capability. To do this, they employed the CFAR100 dataset and created a discontinuous metric that "measures a network's ability to reconstruct a dataset as the average number of test data with squared reconstruction error below cutoff." Results showed that in certain cases using their reconstruction metric, they saw "sharp and unpredictable image reconstruction ability that qualitatively matches published emergent abilities"[1].

To conclude, the authors acknowledge that their tests are not completely similar to the original paper as some data wasn't available to replicate the original tests. However, the introduction of this hypothesis and the supporting evidence

provided indicate the need for further evaluation and caution with claims of emergent capabilities in Large Language Models.

References

[1] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. "Are emergent abilities of Large Language Models a mirage?" In: *arXiv preprint arXiv:2304.15004* (2023).

[2] Jason Wei et al. *Emergent Abilities of Large Language Models*. 2022. arXiv: `2206.07682 [cs.CL]`.

[3] Aarohi Srivastava et al. "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models". In: *arXiv preprint arXiv:2206.04615* (2022).

[4] URL: `https://r2r.tech/articles/accuracy-precision-linearity-and-resolution-web-guiding-understanding-terminology`.