Research Seminar Data Science Task 2 — Paper 18: Are Emergent LLM Abilities a Mirage?

Daan Brugmans (S1080742)

I. Summary

Schaeffer et al. investigate the phenomenon of "emergent abilities", where Large Language Models (LLMs) unexpectedly gain newfound capabilities that make them vastly outperform smaller-scale models. The authors find that, although such emergent abilities are most often attributed to the scale of modern LLMs, they actually arise mostly from the specific metrics used to measure model performance. Specifically, the authors find that emergent abilities only arise for non-linear and discontinuous metrics, and that these abilities disappear when linear and/or continuous metrics are used instead.

II. Evidence

The authors start with a theoretical explanation of how emergent abilities may arise from metric choice, explained using a specific and concrete example. They then perform experiments based on their theory on the InstructGPT and GPT-3 models, and find that their hypothesis holds in practice: when measured with linear and continuous metrics, GPT models do not suddenly obtain emergent abilities at a certain parameter scale, but they develop them semi-linearly as parameter size increases. The authors perform an experiment where LLMs of different architectural families are benchmarked on BIG-Bench, a collection of benchmarking metrics for LLMs. They find that within the 39 benchmarks that comprise BIG-Bench, only 5 imply the presence of emergent abilities, and 2 of those account over 90% of emergent ability claims (Exact String Match and Multiple Choice Grade). These five benchmarks are all non-linear and/or discontinuous. Linear and continuous metrics of BIG-Bench show no emergent abilities. Finally, the authors perform an experiment in which they artificially design a discontinuous metric such that it appears to show emergent abilities for an autoencoder with only one hidden layer.

III. Strengths

- Schaeffer et al.'s findings seem quite impactful: emergent abilities of LLMs are a generally known (if not accepted) phenomenon by many practitioners of data science, and the authors show that that may be unfounded, which has significant implications for future LLMs' capabilities.
- By performing their experiments on InstructGPT and GPT-3, which are publicly available via API, the authors' work is more reproducible.
- The authors made good use of visualizations, which make the presence and lack of emergent abilities in measured performance clear.
- In the appendices, the authors elaborate further on some of the issues with non-linear and discontinuous metrics.

IV. Weaknesses

- The authors do not provide a publicly available source of code used to calculate their quantitative findings.

V. Evaluation

I would recommend acceptance.

VI. Quality of Writing

The paper is a good read. The authors make a successful attempt at explaining, both mathematically and in natural language, how emergent abilities can arise from the metrics chosen. They do so in an easy-to-follow manner.

VII. Queries

- In their discussion, the authors state that they "[...] emphasize that nothing in this paper should be interpreted as claiming that large language models cannot display emergent abilities; rather, our message is that previously claimed emergent abilities [...] might likely be a mirage induced by researcher analyses". As presenters/reviewers of this paper, in your opinion, how can researchers make sure that any perceived emergent abilities can be confirmed or denied with solid certainty, assuming that the findings of this paper hold true?