Assessing the performance of LLMs: the problem of emergent abilities

Céline Budding
Philosophy & Ethics group
Eindhoven University of Technology, The Netherlands
c.e.budding@tue.nl

Abstract (1 page)

Following the Trilogues, large language models (LLMs)–as a type of general-purpose AI system–are to be covered under the AI Act (European Parliament, 2023). Given the widespread use of LLMs, it seems particularly pertinent to ensure that these systems are reliable and trustworthy, but it is still unclear how to accurately assess their performance. In this contribution, I highlight a specific challenge that arises in the context of LLMs: the assessment of whether a system acquires so-called emergent abilities. Then, I suggest that this requires the development of both conceptual criteria and technical methods, such as interventions, to study the internal causal processing of these systems.

Although deep neural networks and LLMs face some similar challenges, such as the black box problems, LLMs also seem to raise new questions. One example is that LLMs have been argued to develop emergent abilities (Wei et al., 2022): abilities that cannot simply be explained by the scaling of the models. Examples of such emergent abilities are the acquisition of theory of mind (Kosinski, 2023; Ullman, 2023), knowledge (Yildirim and Paul, 2023), or even intelligence (y Arcas, 2022). Although the extent to which LLMs acquire such emergent abilities is a topic of debate (Schaeffer et al., 2023), the question of when to attribute additional abilities deserves attention. On the one hand, it is well-known that humans are prone to anthropomorphism, so undue attribution of emergent abilities like knowledge should be avoided (Shanahan, 2022). On the other hand, however, if LLMs were to acquire abilities beyond mere next-word prediction, it is important to accurately asses this.

Solving this problem requires both conceptual and technical innovations. Currently, attribution of emergent abilities is frequently based on observed behavior or interaction with an LLM. Humans are, however, prone to anthropomorphism and might be unable to accurately assess the actual competence of a system (Shanahan, 2022). Therefore, there is a need for standardized evaluation practices, for example in the form of clear conceptual criteria for e.g. knowledge and understanding. Ideally, these criteria would encompass both the behavior of the system, as well as their internal processing. Such criteria would both clarify what exact abilities an LLM is thought to acquire, and provide measurable criteria to evaluate this.

The evaluation of such criteria also requires technical innovations, in particular insight into the internal causal processing of LLMs. This is a notoriously complex problem due to the opacity of these systems, however, so novel approaches might be needed. One approach that deserves further attention are intervention methods (e.g. Meng et al., 2022), which aim to find small edits to the internal parameters of the network such that the behavior is changed in a predictable way. While the main goal of these methods is to find ways to more easily control and fix the behavior of large-scale systems, interventions can also be used to probe and test the internal processing of the system, by systematically changing the internal parameters and observing the resulting effects. In this way, interventions can be used to identify and localize causal processes within the network itself, thereby providing the necessary information for evaluation of the abilities of the system.

Overall, the goals of this contribution are threefold: 1) to argue that potential emergent abilities in LLMs raise new questions for standardization, 2) that both clear conceptual criteria and an understanding of the internal processing of LLMs is necessary to accurately evaluate their performance, and 3) that interventions are a promising method to gain such insights and therefore deserve further scrutiny, both in the context of standardization and the technical literature.

References

European Parliament. 2023. Artificial intelligence act: deal on comprehensive rules for trustworthy ai. https://europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*.

Murray Shanahan. 2022. Talking about large language models. *arXiv preprint arXiv:2212.03551*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Blaise Agüera y Arcas. 2022. Do large language models understand us? *Daedalus*, 151(2):183–197.

Ilker Yildirim and LA Paul. 2023. From task structures to world models: What do llms know? *arXiv preprint arXiv:2310.04276*.