Explanability in the Era of Large language models

Hengli Li
Yuanpei College
Peking University
2000017754@stu.pku.edu.cn

Abstract

The advent of Large Language Models (LLMs) has revolutionized natural language processing, exemplified by models like ChatGPT, llama2, and Gemini. This paper explores the capabilities and challenges posed by LLMs, considering the implications of the scaling law and recent investigations into emergent abilities. Investigating explainability in LLMs, we delve into research by Schaeffer et al. and considerations from "On the Dangers of Stochastic Parrots," highlighting biases, static dataset constraints, and interpretability issues. While LLMs excel in benchmarks, questions persist about their real-world efficacy. We advocate for interdisciplinary collaboration, emphasizing advancements in model interpretability, fairness, and responsible deployment for unlocking the full potential of LLMs.

1 Introduction

Since the advent of ChatGPT, a multitude of distinguished language models has been developed by diverse entities. One noteworthy exemplar is llama2 [7], a widely employed open-source large language model. More recently, Google introduced Gemini. These models undergo extensive training employing voluminous datasets, thereby conferring upon them formidable capabilities in language comprehension and overall proficiency. Nevertheless, the capacities inherent in Large Language Models (LLMs) persist as an enigma, with their limitations yet to be comprehensively delineated. The uncertainties surrounding the scope of these models' potential accomplishments and constraints constitute an ongoing area of inquiry. An influential empirical tenet [3] governing LLMs is encapsulated in the scaling law, positing that the expansion in the size of these models may engender unforeseen capabilities. As illustrated in Figure 1, various assessments of LLMs manifest a sudden surge in metric scores, a phenomenon that has engendered astonishment within the research community. This development precipitates probing inquiries into the learning mechanisms of LLMs, the underlying causative factors of these enigmatic phenomena, and the stimuli precipitating sudden augmentations in their capabilities.

2 Explainability in the era of LLM

An understanding of LLMs is intricately interwoven with the domain of Explainable Artificial Intelligence, assuming a pivotal role in our capacity to regulate these models and cultivate trust between human operators and automated systems. Researchers are actively engaged in the pursuit of answers to these imperative inquiries. Particularly notable amidst recent investigative efforts is the work undertaken by Schaeffer and colleagues, as delineated in the publication entitled "Are Emergent Abilities of Large Language Models a Mirage?" [5]. Their examination delves into the refinement of large language models through a reevaluation of the scaling law. They posit that the predictability of the abilities inherent in large language models may not be entirely inscrutable when evaluated under specific metrics. Their findings contest the supposition that the scaling law might be a mere illusion, propounding that "emergent abilities" could be influenced by three primary factors: non-linearity in per-token error rate metrics, inadequacies in data samples for testing, and a paucity of large language models utilized in the testing phase [5].

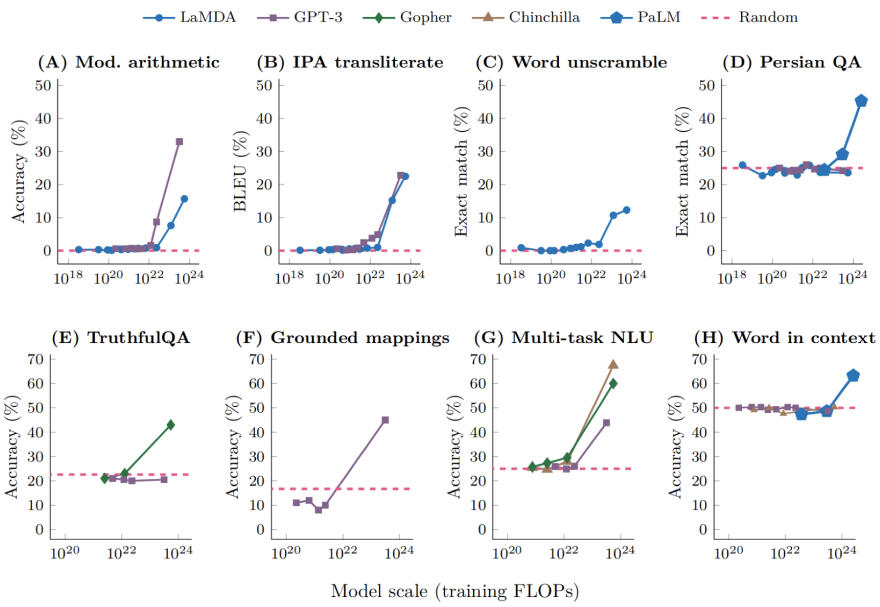

Figure 1: Scaling Law [3]

Conversely, the scholarly endeavor titled "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" [1] not only explores the prospective perils linked with expansive language models but also provides significant elucidations about their behavioral characteristics. The authors present a viewpoint that challenges the prevalent notion that larger models invariably correspond to enhanced performance. A pivotal insight brought to the fore is that the sheer magnitude of a language model does not guarantee inclusivity [1]. This underscores the critical issue of the overrepresentation of particular minor viewpoints, giving rise to the inadvertent encoding of bias and potential harms. The source of bias lies in the predominantly online origin of data for Large Language Models (LLMs), wherein the disproportionate representation of younger users, who are adept at articulating their thoughts online, prevails. Furthermore, economic disparities among nations contribute to an imbalance, as individuals in developed countries, endowed with greater internet access due to financial resources, tend to have their opinions overrepresented. In contrast, voices from marginalized populations, such as those from economically deprived areas or older demographics, are frequently marginalized. Despite their limited online presence, it is imperative to acknowledge the value and importance of their perspectives. The exclusion of these viewpoints should not diminish their significance, as they offer unique insights that may elude the notice of current models.

In addition to these challenges, Large Language Models (LLMs) confront the constraint of being trained on static datasets, rendering them immutable. This challenge is compounded by the neural network's susceptibility to catastrophic forgetting [2], a well-documented issue. Consequently, effecting modifications to the parameters of a network or altering the capabilities of a large language model post-training poses formidable challenges. Moreover, prevailing methods for estimating data contribution suffer from elevated computational complexity, rendering their application in the context of large language models impractical. This intricacy hampers our comprehension of the individual impact of each data point, leaving us in the dark about how data contributes to biases and, more crucially, how to effectively mitigate these biases.

The intricate interplay between the static nature of datasets, the challenges inherent in neural network adaptability, and the computational complexities of data contribution estimation [6] underscores the pressing necessity for advancements in these domains. Such advancements are pivotal to enhancing the flexibility, interpretability, and fairness of large language models.

Large language models demonstrate outstanding performance across diverse benchmarks [1], leading many to perceive their capabilities as nearing human-level proficiency. Nonetheless, a critical inquiry arises: does the attainment of exceptional results on benchmarks equate to competence in real-world scenarios? Can the achievement of state-of-the-art outcomes in benchmarks, such as the Stanford Natural Language Inference benchmark [4], genuinely reflect a model's comprehensive grasp of language? These observations give rise to a noteworthy concern, one that may potentially

guide researchers towards misguided research trajectories. Sole reliance on numerical benchmarks may inadvertently foster unwarranted confidence in quantitative metrics, thereby eclipsing pivotal considerations in the development of artificial intelligence. The peril lies in the oversight of nuanced and indispensable components, leading to an incomplete comprehension of the genuine capabilities and constraints inherent in these advanced language models.

3 Conclusion

In conclusion, the ascendancy of Large Language Models (LLMs) has introduced a spectrum of opportunities and challenges. Although the scaling law implies potential emergent capabilities with augmented model size, contemporary research challenges the predictability of these phenomena. Delving into the intricacies of LLMs, encompassing biases, reliance on static datasets, and interpretability issues, has assumed a paramount significance. Despite their exceptional performance on benchmarks, apprehensions persist regarding the practical utility of these models in real-world contexts. Looking ahead, fostering interdisciplinary collaboration and advancing interpretability and fairness are imperative steps to responsibly unlock the complete potential of LLMs.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922. 2

[2] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3:128–135, 1999. URL https://api.semanticscholar.org/CorpusID:2691726. 2

[3] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. 1, 2

[4] Bill MacCartney and Christopher D. Manning. Modeling semantic containment and exclusion in natural language inference. In Donia Scott and Hans Uszkoreit, editors, *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL https://aclanthology.org/C08-1066. 2

[5] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage?, 2023. 1

[6] Stephanie Schoch, Ritwick Mishra, and Yangfeng Ji. Data selection for fine-tuning large language models using transferred shapley values, 2023. 2

[7] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 1