Do Large Language Models Learn to Human-Like Learn?

Jesse Roberts

Vanderbilt University
jesse.roberts@vanderbilt.edu

Abstract

Human-like learning refers to the learning done in the lifetime of the individual. However, the architecture of the human brain has been developed over millennia and represents a long process of evolutionary learning which could be viewed as a form of pre-training. Large language models (LLMs), after pre-training on large amounts of data, exhibit a form of learning referred to as in-context learning (ICL). Consistent with human-like learning, LLMs are able to use ICL to perform novel tasks with few examples and to interpret the examples through the lens of their prior experience.

Introduction

Transformer based neural networks have led to a number of recent advances in natural language processing and inference (Vaswani et al. 2017). These large language models (LLMs) acquire remarkable abilities through a form of unsupervised learning in which part of the data is hidden and the model is required to reproduce it, a form of cloze task which is similar to de-noising. The model parameters are updated to improve performance on this and similar *pre-training* tasks.

The number of examples required to achieve a language processing ability similar to a child is roughly six orders of magnitude larger than that required by a human (Warstadt et al. 2023). This pre-training process bears little resemblance to the behaviors identified as consistent with human-like learning (Langley 2022). Further, current language model architectures are incapable of adjusting their architectures or parameters based on interactions and are, in this way, incapable of learning.

Through pre-training, language models acquire an alternative means of learning referred to as in-context learning (ICL) which is unique in connectionist literature as it does not involve altering parameters and may not experience *catastrophic forgetting*, though forgetting does occur (Coleman, Hurtado, and Lomonaco 2023). Robustly establishing the presence (or absence) of catastrophic forgetting effects in ICL is a target of my future work.

Through ICL, LLMs perform tasks for which they have little relevant pre-training given a small number of examples (Radford et al. 2019). The examples are interpreted based on the LLM's pre-training and prior interactions. So, while pre-training is not human-like, language models clearly exhibit facets of human-like learning through ICL (Langley 2022).

Emergent Human-Like Learning

A model which acquires the ability to human-like learn through a lengthy pre-training process is consistent with the development of human-like learning in humans.

The human brain is not developed in each individual. Rather, the brain's general architecture is inherited and represents countless generations of improvements. The underlying neurological mechanisms which give rise to specific observed behaviors are not understood sufficiently to make a strong claim regarding the provenance of learning. It is reasonable to hypothesize that human-like learning is an ability that has been acquired, at least partially, through a form of evolutionary *pre-training*. Therefore, human-like learning in humans is to some degree an emergent behavior.

Proposed Presentation

In the proposed talk I will (1) propose that humans have not achieved human-like learning absent of a significant pre-training process, (2) provide an analysis of ICL in light of the facets of human-like learning given in (Langley 2022), and (3) identify the facets of human-like learning which have not been sufficiently explored in ICL. These under-explored facets, like incremental effects, constitute an important hole in the current understanding of language model behavior and its relationship to human-like learning.

References

Coleman, E. N.; Hurtado, J.; and Lomonaco, V. 2023. In-context Interference in Chat-based Large Language Models. *arXiv preprint arXiv:2309.12727*.

Langley, P. 2022. The computational gauntlet of human-like learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12268–12273.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Warstadt, A.; Choshen, L.; Mueller, A.; Williams, A.; Wilcox, E.; and Zhuang, C. 2023. Call for Papers–The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.