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. I examine the constraints which typify human-like learning and propose that LLMs may learn to exhibit human-like learning simply by training on human generated text.

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) (Brown et al. 2020) which is unique in connectionist literature as it does not involve altering parameters and therefore 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 an important target for 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, though sufficient ICL examples can override pre-training (Wei et al. 2023). So, while pre-training is not human-like, language models clearly exhibit facets of human-like learning through ICL (Langley 2022).

In the remaining sections of this paper, I (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 constitute an important hole in the current understanding of language model behavior and its relationship to human-like learning.

Emergent Human-Like Learning

I hold that 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. Concepts like cognitive modularity are tied strongly to evolutionary co-development of behavior and hardware (Barrett and Kurzban 2006). The interplay of brain hardware development and behavior modification across a sea of time is believed to have led to the specialization of modular structures. Further, it is established that neuro-typical individuals learn differently as compared to their dyslexic peers (Alsulami 2019), and that those with dyslexia have consistent differences in their brain structures as compared to the neurotypical cohort (Sun, Lee, and Kirby 2010).

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. However, 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*.

What is ICL?

In-context learning (ICL) refers to a model learning to perform a task after being given a single or small number of examples in the model's context. Importantly, ICL does not involve any changes to model weights. So, novel task abilities are necessarily a result of interactions between the tokens in the context and the pretrained model. For a longer review of work regarding ICL refer to (Dong et al. 2022).

As a clarification, not all language models acquire the ability to in-context learn. It has been shown to emerge when data possesses certain properties common in language. When these properties are absent, models perform tasks using stored information in weights but will not improve performance with the presentation of in-context examples (Singh et al. 2023).

Human-Like Learning Constraints

In this section, I consider the constraining attributes of human-like learning presented in (Langley 2022) and examine current empirical and theoretical research helping to establish whether human-like learning associated behaviors have been found to be present in LLMs, specifically when they engage in ICL.

Learning Involves the Acquisition of Modular Cognitive Structures. Many authors have held that cognitive structures in the human mind are modular with any precise meaning of modular being contested, like that in (Fodor 2000) requiring that modules be separated and specific. In (Barrett and Kurzban 2006), the authors provide an empirical view of modularity:

> ...functional magnetic resonance imaging (fMRI) might demonstrate the interaction of multiple systems and use of information from multiple sources, such findings do not falsify a hypothesis of principled and specialized use of information by dedicated systems. Empirically, what counts as evidence for or against a particular hypothesis about modularity turns on having a theory that predicts which inputs are relevant and, therefore, the psychological effects one expects to observe in different situations.

Modularity, in this notion, is not undermined by co-recruitment or distributed processing which is commonly found in fMRI-based human studies. In their view, such evidence simply serves to show that *encapsulation* is not a requisite component of modularity.

Adopting a similar perspective, I propose that LLMs possess functionally modular, though not encapsulated, knowledge structures. In (Bayazit et al. 2023), the authors show that domain specific knowledge sub-networks are identifiable and separable in GPT-2 such that, after ablation, the network is unable to perform related tasks but maintains unrelated knowledge and language ability. So, even though the entire network is executed for any task, it seems only a relatively small, modular subnet constitutes the pertinent knowledge for the task.

That being said, this does not suggest that structures are acquired during ICL since the network weights aren't being changed.

When a token is placed in the context of a transformer, three learned linear transformations are applied. The key and query transforms provide a representation that is used to find the attention weight placed on each token. Then, for each token in the context, $t_i \in S$, the associated attention, α_i, and value transform is used to create an admixture, $\Sigma_i \alpha_i \cdot V(t_i)$. It may be said that the unpacked value representations are structures acquired through ICL.

Learned Cognitive Structures Can be Composed During Performance. Given the above notion of modularity, a cognitive structure within the network may be activated by a token in the ICL prompt. However, by having attention spread across multiple tokens, the output is generated from the compositions of individual modular structures (tokens). Each of these tokens then becomes a query that is used to create a set of contextually based representations. These representations are themselves composed into a single representation over the context given the query. The subsequent layers perform the same set of actions, resulting in compositions of compositions.

It is important to note that, while this can serve to create powerful compositions, transformers are not capable of arbitrary composition (Roberts 2023) without recursion. Though models like the decoder-only transformer are capable of recursion, language models don't typically learn this behavior as evidenced by the need to explicitly illicit recursive problem solving behavior through chain of thought prompting (CoT) (Wei et al. 2022).

Many Learned Cognitive Structures Are Relational. It is well established that language models based on the transformer architecture learn relational information (Rezaee and Camacho-Collados 2022; Bouraoui, Camacho-Collados, and Schockaert 2020; Petroni et al. 2019). However, the relevant question is, do language models learn novel relational structures through ICL?

It has been shown that ICL facilitates learning truly new relational information (Kossen, Gal, and Rainforth 2024). However, this does not suggest that ICL permits the induction of novel types of relational structure as is necessary when prompted with semantically unrelated labels (SUL-ICL). This ability has been shown to tend to emerge when language models are massively scaled (Wei et al. 2023) like in the case of PaLM-50B (Anil et al. 2023).

Expertise Is Acquired In a Piecemeal Manner. As discussed, each individual token presented through ICL results in an additional modular, composable structure which, by nature, is acquired in a piecemeal manner. However, to some relevance here, the study of human behavior has revealed many distinguishing facets present in expert behavior, like the use of heuristics as opposed to a reliance on rules, which are absent in the novice (Palmeri and Cottrell 2010). So, a more nuanced question may be, do language models in-context learn expert-like performance and behavior? A review of the current literature regarding ICL **suggests this has not been addressed**.

In (Anderson 1995), a link between long term mem-

ory and expert behavior is established. I recommend future work should investigate the effects of ICL on language model long-term working memory (LTWM) (Sohn and Doane 2003) for items of the type presented in prompting, to empirically establish the relationship of ICL and expert behavior.

Learning Is An Incremental Activity That Processes One Experience At a Time. The work in (Kossen, Gal, and Rainforth 2024) shows that ICL permits a language model to develop improved task performance with each in-context example. However, in most empirical work on ICL the test method presents all in-context examples as a small batch as opposed to interleaved experience and inference as may often be the case in human interaction. While interleaved example and inference may be a common practical prompt pattern in language model use, a review of the literature suggests this its effects on ICL performance **have not been explicitly considered**. However, research suggesting ICL suffers from a form of forgetting (Coleman, Hurtado, and Lomonaco 2023) suggests that interleaved ICL may have a mitigated effect.

Learning Is Guided by Prior Experience. In (Kossen, Gal, and Rainforth 2024), the authors show that providing a single incorrect example followed by correct examples harms the model's performance until correct in-context examples sufficiently outnumber the incorrect example. This shows that learning is guided by the prior experience to an extent.

However, in (Langley 2022) the motivational examples call for a more significant treatment of this question. At the time of writing, no work was identified in the literature that explicitly considered the degree to which subsequent ICL examples interfere (constructively or destructively) with prior examples.

Cognitive Structures Are Acquired and Refined Rapidly. ICL drew significant attention as a unique ability that language models learn to exhibit. As already described, the hallmark of ICL is the ability to learn novel tasks from one to a few examples. Language models are certainly able to acquire (Radford et al. 2019) and refine (Kossen, Gal, and Rainforth 2024) knowledge structures rapidly with few examples through ICL.

Conclusions

ICL is a powerful and unique emergent ability present in certain language models of sufficient size. When the pretraining of the language model is seen analogically as a counterpart to the evolution of the human brain, ICL stands as a reasonable counterpart to human-like learning in language models. I have examined the constraints defined in (Langley 2022) and applied the resulting insightful lens to ICL in language models by examining the literature and identifying the challenges within the gauntlet of human-like learning already met by ICL and those that stand as important future work.

The development of expertise and the effect of incremental experience have not been sufficiently considered in the literature. Further, the composition of transformers is bounded by the depth of the model given most models are unable to engage in arbitrary recursion. However, all other constraints given in the motivating paper have either been shown to be empirically or theoretically met by ICL.

References

Alsulami, S. G. 2019. The Role of Memory in Dyslexia. *International Journal of Education and Literacy Studies*, 7(4): 1–7.

Anderson, J. R. 1995. Cognitive psychology and its implications (4" ed.). *Now York: WH Freeman and Company*.

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Barrett, H. C.; and Kurzban, R. 2006. Modularity in cognition: framing the debate. *Psychological review*, 113(3): 628.

Bayazit, D.; Foroutan, N.; Chen, Z.; Weiss, G.; and Bosselut, A. 2023. Discovering knowledge-critical subnetworks in pretrained language models. *arXiv preprint arXiv:2310.03084*.

Bouraoui, Z.; Camacho-Collados, J.; and Schockaert, S. 2020. Inducing relational knowledge from BERT. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7456–7463.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

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

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; and Sui, Z. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Fodor, J. A. 2000. *The mind doesn't work that way: The scope and limits of computational psychology*. MIT press.

Kossen, J.; Gal, Y.; and Rainforth, T. 2024. In-Context Learning Learns Label Relationships but Is Not Conventional Learning. In *The Twelfth International Conference on Learning Representations*.

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

Palmeri, T. J.; and Cottrell, G. W. 2010. Modeling perceptual expertise.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

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.

Rezaee, K.; and Camacho-Collados, J. 2022. Probing Relational Knowledge in Language Models via Word Analogies. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 3930–3936.

Roberts, J. 2023. On the Computational Power of Decoder-Only Transformer Language Models. *arXiv preprint arXiv:2305.17026*.

Singh, A. K.; Chan, S. C.; Moskovitz, T.; Grant, E.; Saxe, A. M.; and Hill, F. 2023. The Transient Nature of Emergent In-Context Learning in Transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sohn, Y. W.; and Doane, S. M. 2003. Roles of working memory capacity and long-term working memory skill in complex task performance. *Memory & cognition*, 31: 458–466.

Sun, Y.-F.; Lee, J.-S.; and Kirby, R. 2010. Brain imaging findings in dyslexia. *Pediatrics & Neonatology*, 51(2): 89–96.

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*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wei, J.; Wei, J.; Tay, Y.; Tran, D.; Webson, A.; Lu, Y.; Chen, X.; Liu, H.; Huang, D.; Zhou, D.; and Ma, T. 2023. Larger language models do in-context learning differently.