On the Ability To Reason and Large Language Models

Bill Cochran

July 21, 2025

1 The Idealized Language Model

We begin our exploration by examining what an idealized neural network can represent. Let $\mathcal{L}_i$ denote the class of i-th order logical predicates, defined inductively as follows.

Definition 1 (Inductive Hierarchy of Logical Predicates)**.** *Let $\mathcal{L}_0$ denote the set of atomic predicates over tokens (e.g., symbol identity, part of speech, punctuation).*

We define the hierarchy of logical predicate classes $\mathcal{L}_n$ inductively as follows:

- ***Base case:*** *$\mathcal{L}_0$ consists of atomic predicates on tokens.*
- ***Inductive step:*** *For $n \geq 0$,*

  $$\mathcal{L}_{n+1} = \{Q\, p \mid p \in \mathcal{L}_n,\, Q \in \{\forall, \exists\}\} \cup \{f(p_1, \ldots, p_k) \mid p_i \in \mathcal{L}_n\}$$

  where f denotes any computable composition function (e.g., logical conjunction, implication, arithmetic).

Let $N_{d,w}$ describe a dense feedforward neural network of depth d with width w at each layer. By dense, we mean every node in layer d is connected to every node in $d-1$ and $d+1$.

We first demonstrate that for all $w \in \mathbb{N}$, there exists a predicate $p \in \mathcal{L}_2$ that cannot be represented by any one-layer network $N_{1,w}$. In particular, consider the predicate:

"There are at least $w + 1$ true propositions in the input."

This predicate is clearly in $\mathcal{L}_2$: it quantifies over a set of atomic (i.e., $\mathcal{L}_0$) truth values and performs a cardinality comparison. It cannot be represented by $N_{1,w}$, since any one-layer network of width w can at most linearly combine w truth signals. This fails to capture any threshold predicate over more than w inputs — such predicates lie in the network's null space.

It is therefore trivial to observe that we can construct a countably infinite set of such unrepresentable predicates for each fixed w.

Now suppose, inductively, that given a network $N_{d,w}$, we can construct a countably infinite set of predicates $P_{d+1} \subset \mathcal{L}_{d+1}$ that require depth $d + 1$ for representation. Then we can construct a countably infinite set $P_{d+2} \subset \mathcal{L}_{d+2}$ by applying a second-order quantifier to P_{d+1}:

"There are at least $d \cdot w + 1$ true predicates in P_{d+1}."

This predicate belongs to $\mathcal{L}_{d+2}$ and, by the same argument, cannot be represented by any network of depth $d + 1$ and width w. The construction proceeds inductively, bounding the logical expressiveness of $N_{d,w}$ strictly below $\mathcal{L}_{d+1}$. In other words, this statement is guaranteed to produce an hallucination.

These predicates use a very simple principle to exhaust the logical capability of the any network. By asking the network to count one more extremely complex object than it can possibly model ensures that the model will hallucinate because it just doesn't have the entropy necessary to dinguish that many concepts.