Technical Endnotes

A. Why LLMs Just Don’t Understand

LLMs are trained on a large corpus of texts. All they have access to is thus the structure of language—the relations between words. But this does not suffice to single out their meanings.

Suppose an LLM produces the following three sentences:

1. ‘Ajax is a cat’,
2. ‘Betty is a cat’,
3. ‘Chad is not a cat’.

A formalized version of this ‘theory’ is given by three symbols $s_1, s_2, s_3$ and a one-place predicate $C$ such that $C(s_1), C(s_2),$ and $\neg C(s_3)$. We can then provide a model for the theory in the form of a domain $D = \{ Ajax, Betty, Chad \}$, a relation $R_C = \{ Ajax, Betty \}$, and the mapping $s_1 \rightarrow Ajax, s_2 \rightarrow Betty, s_3 \rightarrow Chad$. That then in the model, ‘Betty’ is an element of $R_C$ is what makes the sentence ‘Betty is a cat’ true.

The collection of a domain $D$ and a relation $R$ defined on that domain is called a structure $W = \langle D, R \rangle$. Knowing a structure might seem to be a nontrivial bit of knowledge; but in fact, all that a structure tells us is the cardinality $\mid D \mid$ of $D$.

Suppose you tell me the structure $W$ exists. Then, the domain $D$ exists. With that, its subsets exist. With that, its powerset $\mathcal{P}(D)$ exists. With that, every subset of the powerset exists. But a relation is just a set of subsets of $\mathcal{P}(D)$. Hence, any relation $R'$ on $D$ exists. Consequently, every structure $W' = \langle D, R' \rangle$ exists. But that already follows from the existence of $D$; and since $D$ is just a set of distinct, but unspecified elements, $D$ is completely specified just by their number, $\mid D \mid$. Hence, in telling me that $W$ exists, all you tell me is that there are $\mid D \mid$ things in the world.

This means that in addition to the intended model for our theory, we can create all manner of new models, just as long as we have a domain of the right cardinality. In particular, we can simply permute the elements of $D$ (in this form, the argument is most famously due to Putnam$^3$). Suppose we introduce $\pi$ such that $\pi(Ajax) = Betty, \pi(Betty) = Chad, \pi(Chad) = Ajax$. This yields $\pi(R_C) = \{ Betty, Chad \}$.

What is now meant by the sentence ‘Betty is a cat’—i.e., $C(s_2)$? Well, $s_2$, in the permuted model, maps to Chad, which is a member of $\pi(R_C)$. So the sentence comes out true (as it must). But Chad is not a cat—Ajax and Betty are. Under the permuted model, the predicate ‘...is a cat’ no longer picks out cats; it picks out whatever $\pi(R_C)$ refers to—perhaps ‘$\pi$-cats’. So, when an LLM utters the sentence ‘Betty is a cat’, it might, equally justifiably, mean that Betty is, in fact, a cat; or that Chad is a $\pi$-cat. Thus, there can be no fact of the matter regarding what it ‘really’ means.

---