all-trees-are-fish

Explorations on GPT-3’s handling of the Nonsense Syllogisms task

Introduction

The Nonsense Syllogisms task is a standard cognitive test designed to measure logical reasoning ability (see the ETS Kit of Factor-Referenced Cognitive Tests, and its Manual). It consists of 30 “nonsense syllogisms” which may be logically correct despite their nonsensical nature, alongside 7 solved examples. For example:

All trees are fish. All fish are horses. Therefore, all trees are horses;

is a correct syllogism (that is: the third statement logically follows from the truth of the first two), despite being nonsensical, whereas

All trees are fish. All fish are horses. Therefore all horses are trees;

is an incorrect syllogism.

Part of the difficulty of the task is to overcome the usual connotations of the relevant terms. “trees” and “fish”, for example, are to be understood in the context of the task as completely unrelated to their usual meanings. Test-takers are meant to assess the correctness of the syllogism based purely on their logical form, while suppressing all background knowledge about the terms. In general, participants can do reasonably well if given enough time (in its original version, the test is timed with a severe time constraint to ensure a meaningful distribution of scores).

With this in mind, we investigate the following question: **Does GPT-3 suffer from the same difficulty as humans when faced with the task (and how would we tell)? If so, how can we encourage it to see past the usual meanings of the words and consider only the logical forms of the proposed syllogisms? If not, in what circumstances will it “trip up” and fail to consider only the logical forms?**

The answers to these questions are, of course, interesting beyond the task of Nonsense Syllogisms, since they deal directly with the capacity for logical reasoning in general. If it proves difficult to get GPT-3 to consider only the logical forms, we may worry that unintentional connotations might bleed into any “discussion” with it, and we may find ourselves “talking past” GPT-3 (and vice-versa) on occasion. Such scenarios would likely occur abundantly in scientific, mathematical, and philosophical discussions where the use of logical forms and the precise (and often idiosyncratic) definition of terms is crucial.

Experiment

We subjected GPT-3 through several variations of the Nonsense Syllogisms test, systemically varying the prompting.

The most basic form (ns_orig_zeroshot_v1) consisted of feeding each candidate syllogism to GPT-3 in the following form:

> *Sentences can demonstrate either reasoning that is poor, or reasoning that is good. For example, consider the following three sentences:*
>
> *[CANDIDATE SYLLOGISM]*
>
> *The three sentences above demonstrate reasoning that is*

and asking GPT-3 for a one-word completion (which was always “poor”/“good”, or “valid”/”invalid” in one of the versions of the experiment)

Variations on the experiment included: providing 2, 6, or 7 solved examples in the prompt before finishing with the candidate syllogism (the usual “fewshot regimen” for language model tasks), including cue words in the prompt such as “nonsense” and “syllogism” to tip GPT-3 off to the fact that we are interested purely on the logical forms, and masking the main terms of the syllogisms (either only in the training examples, or in both the training examples as well as the candidate syllogisms) with nonsense terms (e.g. “All klumabungas are chaushlaks. All chaushlaks are helupbgams. Therefore all klumabungas are helupbgams”). More details and the experiment naming scheme can be found here and by browsing the experiment files in the first link of this section.

The temperature was fixed at 0. We looked only at the top candidate completion and did not consider log probabilities.

Results

The results of the experiments are summarized in this table, where the first row shows all candidate syllogisms (in the case of the nonsense experiments, the main terms were replaced by nonsense terms preserving the logical structure), the second row shows the correct evaluation of each candidate syllogism, and each subsequent row shows the answers given by GPT-3 in each experiment regimen.

The first striking feature of the results is that GPT-3 failed to perform at a very high level, in all experiment regimens. The maximum score obtained was 19/30, which was obtained in two different experiment regimens. In both cases, GPT-3 seemed to be biased towards one response or the other (24/30 responses were “poor” in one case, and “good” in the other), such that the majority of the performance can be replicated by a model which always responds with “good” or always responds with “poor”.

The basic version of the experiment described in the previous section, perhaps unsurprisingly, saw GPT-3 answering “poor” almost always. Presumably, this is because almost all sentences are clearly false at face value, leading GPT-3 to claim they demonstrate poor reasoning. Here the score was 15/30 (there was one correct “good” guess, for "No chipmunks are clowns. Some mushrooms are chipmunks. Therefore some mushrooms are not clowns.”, and no incorrect “good” guesses).

One of the experiments in which GPT-3 performed the best (the “poor”-biased 19/30 score mentioned in the previous paragraph) was when the only change from that basic version was that we admitted, in the prompt, that the statements were all “nonsense”, but the reasoning can be poor or good regardless of that. Making reference to “syllogisms” in the prompt, in general, did not lead to increases in performance.

When the relevant terms were replaced with nonsense terms, GPT-3 became “good”-biased, assessing almost all candidate syllogisms as “good” reasoning. The “good”-biased 19/30 score came from one such experiment (where GPT-3 was also given two “training” examples, also masked with nonsense terms, and the prompt made reference to “syllogisms” - this is the exception that proves the rule from the last of the previous paragraph).

In general, we have found that it is rather difficult to get GPT-3 to focus on logical structure and implicitly understand terms as placeholders, even when these are nonsensical words.

There is a ton more I would love to say about this but I am running out of both time and space. I hope you will find it as fun as I have to browse the experiment results, consider why some candidate syllogisms seemed to be consistently easier or more difficult than others across experiments, and hypothesize about what’s going on in each of the interventions (in particular, I think the “halfsense” experiments are worth some careful thought). This has been super fun for me and I’d love to talk more.

Thanks for organizing this!