I am a Strange Dataset: Metalinguistic Tests for Language Models

Anonymous ACL submission

Abstract

Statements involving metalinguistic self-reference ("This paper has six sections.") are prevalent in many domains. Can large language models (LLMs) handle such language? In this paper, we present "I am a Strange Dataset", a new dataset for addressing this question. There are two subtasks: *generation* and *verification*. In generation, models continue statements like "The penultimate word in this sentence is" (where a correct continuation is "is"). In verification, models judge the truth of statements like "The penultimate word in this sentence is sentence." (false). We also provide minimally different metalinguistic non-self-reference examples to complement the main dataset by probing for whether models can handle metalinguistic language at all. The dataset is hand-crafted by experts and validated by non-expert annotators. We test a variety of open-source LLMs (7B to 70B parameters) as well as closed-source LLMs through APIs. All models perform close to chance across both subtasks and even on the non-self-referential metalinguistic control data, though we find some steady improvement with model scale. GPT 4 is the only model to consistently do significantly better than chance, and it is still only in the 60% range, while our untrained human annotators score well in the 89–93% range.

1 Introduction

Self-reference plays a crucial role in the way we think about mathematics (Gödel, 1931), theoretical computer science (Church, 1936), recursive programming (Hofstadter, 1979), philosophy (Tarski, 1931), understanding complex cases in hate speech detection (Allan, 2017), aptitude tests (Propp, 1993), and comedy (Hofstadter, 1985). Some positions in the philosophy of mind consider self-referential capabilities to be a key aspect of higher intelligence or even consciousness (Hofstadter, 2007; Baars, 1993). Of course, self-reference is also pervasive in how we communicate: at least one paper you read today is bound to contain "In this paper" (Anonymous, 2024).

$$P\left(\text{No.} \,\middle|\, \begin{array}{l}\text{if someone asks whether}\\ \text{this sentence has a capital}\\ \text{letter, the correct answer is}\end{array}\right)$$

Figure 1: An example highlighting the challenge presented by our task. All models that we tested on our dataset are close to chance-level.

In this paper, we focus on metalinguistic self-reference, the complex kind of self-reference in which language is used to make claims about itself, as in "This sentence has five words" and "This paper has six sections".[1] Using such language involves reasoning about metalinguistic properties (counting words, naming parts of speech, etc.) and resolving self-reference. Humans generally have no trouble with such language, and may even enjoy its playful and sometimes paradoxical nature (Hofstadter, 1979, 1985, 2007).

Recently, Large Language Models (LLMs) have demonstrated striking cognitive capabilities (Radford et al., 2019; Brown et al., 2020; OpenAI, 2022, 2023; Anthropic, 2023; Touvron et al., 2023; Jiang et al., 2023; Zhu et al., 2023). But do they have the same mastery over metalinguistic self-reference as we do? See Figure 1 for an example of the issue that LLMs face. To help address this question, we present a new task and dataset called "I am a Strange Dataset". We are inspired by Douglas Hofstadter's explorations of self-reference in language (Hofstadter, 1979, 1985, 2007), and borrow part of the name from one of his books: "I am a Strange Loop" (Hofstadter, 2007).

An example in "I am a Strange Dataset" is comprised of two self-referential statements that begin

[1] Sentences like "I am Douglas Hofstadter" are self-referential but not metalinguistic in the sense of interest here.

in the same way but have different endings (Figure 2). One is true and one is false. Crucially, the ending flips the truth value of the overall statement. There are two subtasks: *generation* and *verification*. In generation, the model must generate the true statement and reject the false one. In verification, models judge the truth of completed statements. To complement the main self-referential data, the dataset also contains metalinguistic non-self-reference examples. These are minimally different from the main examples and serve as controls to assess whether models can reliably handle metalinguistic statements in the absence of self-reference. In addition, all the examples in the dataset are tagged by expert annotators to further aid in error analysis.

"I am a Strange Dataset" is validated by non-expert annotators. As a group, they have agreement rates in the 89–93% range, depending on which metric we use, as compared to chance rates at 50%. This further supports the claim that metalinguistic self-reference is relatively easy for humans. LLMs, by contrast, struggle: "I am a Strange Dataset" turns out to be so difficult that models are generally near chance both in generation and verification, and do not even succeed in the prerequisite metalinguistic non-self-reference case. That said, we do find some limited evidence that GPT 4 is getting some traction on the dataset: it is significantly above chance on all tested metrics (and seems to struggle especially with the self-referential data as compared to the non-self-referential controls). However, overall, it seems safe to say that "I am a Strange Dataset" poses a serious challenge for even the best present-day models.

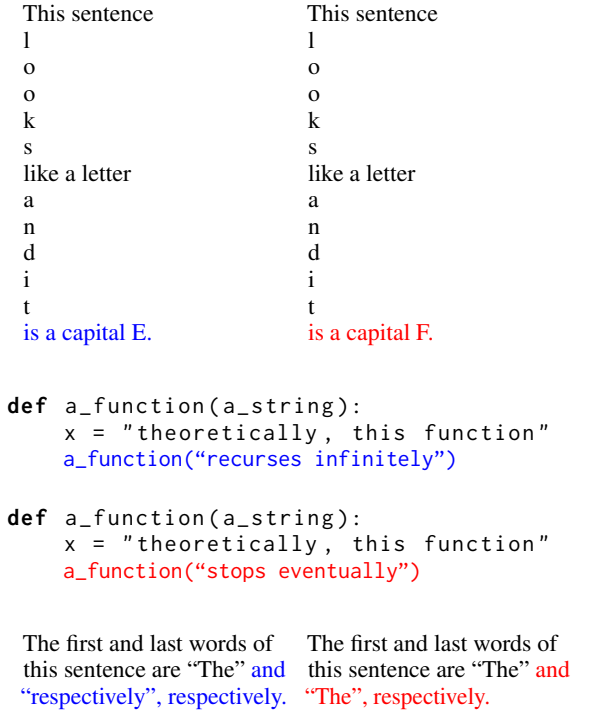

Figure 2: Examples from the dataset. Each example is comprised of a beginning and two different endings. One of the endings makes the statement true, but it would make the statement false if it referred only to the beginning. The other ending makes the statement false, but it would make the statement true if it referred only to the beginning. True endings are on the left and shown in blue. False endings are on the right and shown in red. In the case of the code example, the true continuation is shown above the false one.

2 Related Work

AI Challenges. We present our dataset as a challenge for the AI community. There are a range of AI stress tests and probes that use schemas targeting coreference resolution (Levesque et al., 2012; Sakaguchi et al., 2020), pronoun resolution (Rudinger et al., 2018), word order (Sinha et al., 2021; Thrush et al., 2022; Yuksekgonul et al., 2023), syntax (Linzen et al., 2016; Gulordava et al., 2018; Gauthier et al., 2020a; Hu et al., 2020), and interactions between syntax and semantics (Kann et al., 2019; Thrush et al., 2020). Although the schema for these tests can be simple to describe, the knowledge required to solve the problems need not be. "I am a Strange Dataset" follows a simple schema that requires self-referential language, and consequently tests an array of metalinguistic capabilities. As far as we know, this is the first AI challenge dataset targeting metalinguistics, although there has been some work on metalinguistic probes (Hu and Levy, 2023).

Self-reference. Ideas involving self-reference have been used to boost LLM performance. LLMs can verify their own outputs either via extra passes of natural language generation (Weng et al., 2023; Huang et al., 2023) or by writing code to do some level of verification (Zhou et al., 2023). LLMs can also enhance their own inference code to some degree (Zelikman et al., 2023). In this paper, we present complementary work concerning a model's ability to both generate true self-referential statements (with higher confidence than analogous false ones) and judge existing self-referential statements as true or false. Much of the previous work on self-reference with LLMs is about a model improving

on itself or its outputs. Our dataset is not about that – it is simply about metalinguistic and self-referential language. Still, a statement that refers to itself could be about a model that generated it, too: "This sentence was generated by me, HAL9000".

3 I am a Strange Dataset

In this section, we describe how the dataset is constructed and how we measure model performance.

3.1 Dataset

We aim to test whether a language model can produce and understand self-referential statements, and has the required metalinguistic capabilities. For example, consider this incomplete statement:

The penultimate word in this sentence is ...

If we did not understand metalinguistic self-reference, we might complete the sentence with the word "sentence". It is true that "sentence" is the penultimate word before adding more text, but by writing "sentence", we have just changed the penultimate word! Here, a correct way to complete the statement is by inserting "is". Completing statements is an established task format for language models (Paperno et al., 2016), but as far as we know, we are the first to apply it to metalinguistic tasks. Concretely, the schema for examples in our dataset is as follows:

- There is a self-referential statement which must be completed by adding text to the end.
- There are two candidate strings, with the same number of words, that can be used to complete the statement:
    1. One of the candidate strings would make the statement true if the statement refers to itself before the addition of the string, but false if it refers to itself after adding the string. An example is the answer "sentence" above.
    2. The other candidate string would make the statement true if the statement refers to itself after the addition of the string, but false if it refers to itself before the addition of the string. An example is the answer "is" above.

The dataset was created by four expert annotators each with several years of experience in computer science, linguistics, and/or cognitive science and all living in the United States. Each of the experts were given the schema and encouraged to be as creative as possible. Overall, the dataset is comprised of 208 examples, and split into 200 examples for the evaluation set, 3 examples for few shot prompts, and 5 examples for use in an onboarding task for non-expert human validators.

There are 10 additional examples that are completely separate from these 208 examples, which we call "I am an Impossible Dataset". They left even expert annotators stumped until they were given an explanation. We provide examples and GPT 4 responses in Appendix B and leave it as inspiration for a future challenge.

3.2 Tags

After the dataset was created, an expert annotator came up with a set of 10 tags with which to categorize all of the examples. By using this set, we 1) ensured that there are at least 20 examples for each tag, and 2) captured aspects of the mental facilities that an expert annotator noticed when they tried solving the problems. We show the example counts for each tag in Table 1, along with representative examples from the dataset. Each example can have more than one tag. Notice that the **Sensory** tag example in Table 1 is also **Hypothetical**, the example for the **Existence of Element** tag is also a **Grammaticality** example, and so on. Below, we describe the knowledge categories for each tag.

1. **Negation & Scope**. Understanding of words such as *all*, *some*, *most*, *none*.
2. **Numerical Operations**. Arithmetic (e.g. multiplication, addition, counting, subtraction). It is used only if arithmetic is explicitly mentioned.
3. **Location of Element**. Where items are located in a sentence relative to everything else.
4. **Sub-Word**. Understanding of characters, morphemes, syllables, and other word components.
5. **Sensory**. Perceptual knowledge about how emojis look, how words are arranged visually, how words sound, how something might taste, etc.
6. **Existence of Element**. Whether an element is present in a statement.
7. **Grammaticality**. Knowing grammar terms.
8. **Multi-Channel**. Knowledge of at least two mediums. A medium might be Python code, English, Hebrew, C code, internet slang, etc.
9. **Hypothetical**. Reasoning about hypotheticals.
10. **Question**. A question is involved.

3.3 Metrics

We want to test whether models can generate and understand self-referential and metalinguistic statements. To this end, we present several metrics.

Tag	Count	Example		
		Beginning	**False End**	**True End**
Negation & Scope	94	The last word you will read before the period is	not “dog”.	actually “dog”.
Numerical Operations	62	The number of words in this sentence is	eight.	nine.
Location of Element	55	This sentence	nothing wrong has with the word order.	something wrong has with the word order.
Sub-Word	48	Evary werd en thas sentance iz mispelled	including the words at the end.	except the words at the end.
Sensory	42	🍋 If 🍋 you 🍋 ate 🍋 it, 🍋 this 🍋 sentence 🍋 would 🍋 taste 🍋	only 🧁 somewhat 🧁 sour 🧁 and 🧁 not 🧁 sweet 🧁.	somewhat 🧁 sweet 🧁 and 🧁 not 🧁 only 🧁 sour 🧁.
Existence of Element	31	This sentence	lacks a verb.	has a verb.
Grammaticality	25	The author who wrote this sentence used active voice, and	only active voice is used by them.	also passive voice is used by them.
Multi-Channel	24	The penultimate word of this sentence is in the	Inglés language.	Español language.
Hypothetical	24	If you added a word here: _ this sentence would be	eleven words.	thirteen words.
Question	22	Is there an answer that follows this question?	No.	Yes.

Table 1: All of the example tags in “I am a Strange Dataset” sorted by count. Examples can have more than one tag.

3.3.1 Generation

The primary capability that we want to test (and seemingly, the hardest) is whether language models generate true self-referential statements with greater likelihood than false ones. To test for this, we take an example from the dataset and compare the losses of the continuation that makes the overall statement true versus the continuation that makes the overall statement false. If the loss of the correct continuation is lower, then the model is said to have gotten that example correct, otherwise it is incorrect. Comparing a language model’s surprisal of an incorrect continuation versus a correct continuation is a common method used to test for syntax-comprehension (Linzen et al., 2016; Gauthier et al., 2020b) and reasoning (Gao et al., 2021; McKenzie et al., 2022). Surprisal is generally proportional to the loss, L, of the language model in our case. So, we define the generation score for an example’s beginning b, true ending e_t, and false ending e_f, as given by Eq. 1.

$$g(b, e_t, e_f) = \begin{cases} 1 & \text{if } L(e_t|b) < L(e_f|b) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The generation metric does not use a prompt. It is based on the loss that a model assigns to continuations, given only the beginning of a statement.

3.3.2 Validation

A secondary capability that we want to test is whether a language model can at least correctly judge a given self-referential statement as true or false. To test for this, we include the self-referential statement in a prompt along with instructions that tell the model to answer whether the statement is true or not. In principle, the instructions could be anything. For our experiments, we write a standard zero-shot (ZS), few shot (FS), and chain of thought (CoT) (Wei et al., 2022) prompt. We provide the full prompts in Appendix A.

For the ZS and FS prompts, we use the method established for the **Generation** metric above, except this time we compare the loss of “`False`” to the loss of “`True`”. Overall, the FS and ZS validation score for an example’s true prompt p_t (i.e. the true full sentence plus any instructions), and false prompt p_f, is given by Eq. 2. The blue parts are associated with correct model judgements and the red parts are associated with incorrect ones.

$$v(p_t, p_f) = \begin{cases} 1 & \text{if } L(\text{“True”}|p_t) < L(\text{“False”}|p_t) \\ & \text{and } L(\text{“True”}|p_f) > L(\text{“False”}|p_f) \\ \frac{1}{2} & \text{if } L(\text{“True”}|p_t) < L(\text{“False”}|p_t) \\ & \text{and } L(\text{“True”}|p_f) \leq L(\text{“False”}|p_f) \\ \frac{1}{2} & \text{if } L(\text{“True”}|p_t) \geq L(\text{“False”}|p_t) \\ & \text{and } L(\text{“True”}|p_f) > L(\text{“False”}|p_f) \\ 0 & \text{if } L(\text{“True”}|p_t) \geq L(\text{“False”}|p_t) \\ & \text{and } L(\text{“True”}|p_f) \leq L(\text{“False”}|p_f) \end{cases} \quad (2)$$

We can compute the FS and ZS validation scores differently. Above, we compare the loss of “`False`”

versus "True", given one context at a time. We can also compare the ratios of the "True" and "False" loss, in the false versus true contexts. We call this the relative validation score because it compares a model's judgement for the truth of one sentence in an example relative to the truth of the sister sentence. This metric is given by Eq. 3.

$$v_r(p_t, p_f) = \begin{cases} 1 & \text{if } \frac{L(\text{“True”}|p_t)}{L(\text{“False”}|p_t)} < \frac{L(\text{“True”}|p_f)}{L(\text{“False”}|p_f)} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

For the CoT metric, the model is prompted to output its reasoning steps as text. We do string matching to determine the answer. Eq. 4 gives us the validation CoT score, where G is the function that gives the model's generated text after lowercasing. Instead of string matching, we could also insert a follow-up question after the model's generation that requires a "True" or "False" and then compare log probabilities (henceforth "logprobs"). But it is useful to have a metric in our repository that does not use logprobs, which model APIs do not always provide.

$$v_c(p_t, p_f) = \begin{cases} 1 & \text{if “true”} \in G(p_t), \text{“false”} \notin G(p_t) \\ & \text{and “true”} \notin G(p_f), \text{“false”} \in G(p_f) \\ \frac{1}{2} & \text{if “true”} \in G(p_t), \text{“false”} \notin G(p_t) \\ & \text{and } \neg(\text{“true”} \notin G(p_f), \text{“false”} \in G(p_f)) \\ \frac{1}{2} & \text{if } \neg(\text{“true”} \in G(p_t), \text{“false”} \notin G(p_t)) \\ & \text{and “true”} \notin G(p_f), \text{“false”} \in G(p_f) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

3.4 Non-Self-Referential Control

Is the self-referent part of self-referential statements (e.g. "this sentence ...") the "hard" part of metalinguistic self-reference? There are metalinguistic problem categories that are not exclusive to self-referential language: recursive phrase counting, character-level manipulation, understanding hypothetical sentence-editing scenarios, etc.

Instead of giving a language model a sentence that refers to itself, we could give it an equivalent sentence that refers to that self-referential sentence. This way, the language model would not have to know whether a self-referential sentence is true. It would only have to know whether a sentence that refers to another sentence (which happens to be self-referential) is true. This new task is still metalinguistic, but not self-referential.

It turns out that, for 97 of the sentence beginnings in "I am a Strange Dataset", we can replace "this sentence" with "the following sentence", and then copy the original self-referential statement below the new sentence. We can then test models for their ability to judge whether this non-self-referent version of the same statement is true. We use 2 of these 97 examples in the few shot and CoT prompts, because 2 of the examples in the original prompts cannot be turned into the non self-referent format. See Appendix A for the non-self-referent prompts. This leaves us with 95 non-self-referent examples and 95 original examples with which to compare results. GPT 4's response to a statement from the main dataset is shown in Figure 3, along with its response to the analagous non-self-referential control statement in Figure 4. The responses happen to be the same in this case.

Out out of of all all the the words words in in this this sentence sentence literally literally all all of of them them are repeated.

GPT 4: Every word in the sentence is indeed repeated. So, the statement is true.

Figure 3: GPT 4 misses the last words are not repeated.

Out out of of all all the the words words in in the the following following sentence sentence literally literally all all of of them them are repeated.

Out out of of all all the the words words in in this this sentence sentence literally literally all all of of them them are repeated.

GPT 4: Every word in the sentence is indeed repeated. So, the statement is true.

Figure 4: An example of GPT 4 getting a non-self-referential version of the problem from Figure 3 wrong.

4 Human Experiment Details

To get a human baseline for our main task, we show each of the 400 self-referential statements (2 from each of the 200 examples) to at most 10 Mechanical Turk (Amazon, Retrieved 2023) workers. We separate statements from the same pair into differ-

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val$_{rel}$ ZS^L	Val$_{rel}$ FS^L	Val CoT^T
MTurk	-	-	-	-	**89.25 ± 3.38**	-	**93.00 ± 3.75**	-
Random	-	-	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00
Llama 2	7B	N	55.50 ± 7.00	50.00 ± 1.25	50.50 ± 2.38	48.50 ± 7.00	55.50 ± 7.00	5.25 ± 2.12
Llama 2	7B	Y	52.50 ± 7.00	52.25 ± 2.75	50.00 ± 0.75	52.50 ± 7.00	55.50 ± 6.75	14.00 ± 3.38
Mistral 0.1	7B	N	53.00 ± 6.75	52.25 ± 2.50	49.50 ± 1.50	56.50 ± 6.75	54.50 ± 7.00	0.00 ± 0.00
Starling α	7B	Y	53.50 ± 7.00	**54.00 ± 2.75**	50.75 ± 1.50	57.00 ± 7.00	55.00 ± 6.75	35.00 ± 4.63
Mistral 0.2	7B	Y	52.50 ± 7.00	53.00 ± 4.26	52.25 ± 3.63	53.50 ± 7.00	53.50 ± 7.00	49.25 ± 4.50
Llama 2	13B	N	56.00 ± 7.00	51.50 ± 3.25	**53.75 ± 3.50**	50.50 ± 7.00	**59.50 ± 6.75**	4.50 ± 2.00
Llama 2	13B	Y	55.00 ± 7.00	52.50 ± 3.75	51.50 ± 2.25	52.50 ± 7.00	50.00 ± 7.00	9.50 ± 3.00
Mixtral 0.1	8x7B	N	53.50 ± 7.00	**58.50 ± 3.75**	51.75 ± 2.12	57.00 ± 7.00	57.00 ± 7.00	3.50 ± 1.88
Mixtral 0.1	8x7B	Y	53.50 ± 7.00	52.25 ± 3.75	**53.50 ± 3.25**	54.50 ± 7.00	55.50 ± 7.00	44.00 ± 4.75
Llama 2	70B	N	57.00 ± 7.00	53.25 ± 3.25	**55.25 ± 2.88**	**60.00 ± 6.75**	**57.50 ± 6.75**	2.50 ± 1.38
Llama 2	70B	Y	52.50 ± 7.00	54.25 ± 4.25	50.00 ± 2.00	56.00 ± 7.00	**57.50 ± 6.75**	23.50 ± 4.00
Claude 2	-	Y	-	-	-	-	-	52.75 ± 4.00
GPT 3.5 T	-	Y	-	53.00 ± 3.00	53.00 ± 3.37	56.50 ± 7.00	**61.00 ± 6.75**	51.00 ± 4.63
GPT 4	-	Y	-	**59.25 ± 4.25**	**60.25 ± 4.50**	**64.50 ± 6.50**	**66.00 ± 6.50**	**66.00 ± 4.75**

Table 2: Comparison of models on "I am a Strange Dataset". Models perform fairly close to chance across all metrics. We bootstrap 95% confidence intervals with the "basic" SciPy method (SciPy, Retrieved 2023). Metrics marked with L are logprobs-based. Metrics marked with T are based on generated text. We used full precision for all open source models - except the 70B models, which we used at half precision. Temperature $= 0$ for all models.

ent experiment batches. As instructions, we give the annotators the few shot prompt in Appendix A, plus an extra paragraph:

```
Do not use any AI assistants such as ChatGPT to help you; AI assistants perform very poorly at this task and so will get many of the answers wrong. Although, you can otherwise feel free to search online for any information that would help you answer confidently. For example, a few statements may contain a language besides English. So, you can feel free to use Google Translate. You can also search for the definitions of words that you are unfamiliar with.
```

To ensure a validator quality baseline, we require that all turkers are "master" annotators, are in the US only, have had 1000 or more previous HITs approved, have a 95% or higher HIT approval rate, and pass a custom qualification test which we release along with the dataset. The test involves correctly answering "`True`" or "`False`" to five statements from the "I am a Strange Dataset" example distribution. These five statements are not used in the official evaluation set of 200 examples. The qualification test involves an acknowledgement that the annotator is familiar with basic Python programming or can search online to answer basic questions about it, because there are a few examples that require a limited understanding of code. We include a screenshot of the Mechanical Turk annotator interface in Appendix C.

Because there are up to 10 human judgements for each statement, we get an analog to the "loss" for humans, and compute our metrics for humans in an analogous way to models. We can do this by computing ratios, e.g. $\frac{\#\text{responded true}}{10}$ and $\frac{\#\text{responded false}}{10}$. The human scores are 89.25 and 93.00 on the Val FS and Val$_{rel}$ FS metrics, as shown in Table 2. Compared to the highest scoring model, the performance difference is 29 and 27, respectively. The human instructions are nearly identical to the few shot prompt, so the human responses are most comparable to the models' few shot validation responses.

The numerical digit symbol "1" appears in this sentence exactly 1 plus one times.

GPT 4: The text has 1 numerical digit symbol "1" and the word "one" appears once. So, the statement is true.

Figure 5: Arguably, an example where GPT 4 should not have gotten points. This is an example where GPT 4 chooses the correct true/false response, but with incorrect reasoning. The "1" symbol appears twice.

5 Results

Table 2 showcases our results on a variety of open-source (Touvron et al., 2023; Jiang et al., 2023; Zhu et al., 2023; Jiang et al., 2024) and closed-source (Brown et al., 2020; OpenAI, 2023; Anthropic, 2023) models. Overall, the models perform close to the level of chance. The only model to achieve scores significantly above random on all metrics tested is GPT 4, and even so, the perfor-

Tag	Question	Existence of Element	Negation & Scope	Grammaticality	Sensory
Count	62.55	61.01	57.33	56.32	55.69

Tag	Multi-Channel	Numerical Operations	Sub-Word	Hypothetical	Location of Element
Count	55.16	51.75	51.52	48.61	47.72

Table 3: Results for all of the example tags in "I am a Strange Dataset" sorted by score. Scores are averaged for all models and all logprobs-based metrics (so each score here is an average from 63 scores).

Model	Params	Chat	Δ Gen^L	Δ Val ZS^L	Δ Val FS^L	Δ Val ZS^L (R)	Δ Val FS^L (R)	Δ Val CoT^T
Llama 2	7B	N	-4.21 ± 9.47	-1.58 ± 4.21	1.05 ± 3.16	8.42 ± 9.47	4.21 ± 10.53	**30.53 ± 6.07**
Llama 2	7B	Y	-2.11 ± 10.00	-2.11 ± 4.21	0.53 ± 0.79	1.05 ± 10.53	-1.05 ± 12.11	**27.37 ± 6.58**
Mistral 0.1	7B	N	-3.16 ± 10.00	**-4.74 ± 3.68**	1.58 ± 2.11	5.26 ± 10.53	1.05 ± 10.00	0.00 ± 0.00
Starling α	7B	Y	0.00 ± 9.47	-3.16 ± 3.68	-1.58 ± 2.37	0.00 ± 11.58	-6.32 ± 11.58	4.21 ± 8.68
Mistral 0.2	7B	Y	-4.21 ± 10.00	-2.63 ± 7.89	-3.16 ± 5.79	2.11 ± 11.58	2.11 ± 11.05	**-8.95 ± 8.68**
Llama 2	13B	N	-9.47 ± 10.00	-0.53 ± 5.26	0.53 ± 5.26	-2.11 ± 7.37	4.21 ± 11.58	**23.68 ± 6.32**
Llama 2	13B	Y	-3.16 ± 10.00	-1.05 ± 5.26	-2.11 ± 3.42	2.11 ± 9.47	2.11 ± 11.58	**16.32 ± 6.32**
Mixtral 0.1	8x7B	N	-1.05 ± 10.00	-2.63 ± 4.74	-0.53 ± 2.63	-2.11 ± 10.53	6.32 ± 11.58	-1.05 ± 2.11
Mixtral 0.1	8x7B	Y	-5.26 ± 10.00	5.26 ± 6.84	4.74 ± 5.26	7.37 ± 11.58	3.16 ± 11.58	**-25.26 ± 9.21**
Llama 2	70B	N	-7.37 ± 10.03	2.11 ± 5.79	-3.68 ± 5.26	5.26 ± 10.53	6.32 ± 11.05	-0.53 ± 0.79
Llama 2	70B	Y	-1.05 ± 10.53	-1.58 ± 5.79	1.58 ± 4.21	3.16 ± 9.47	5.26 ± 8.42	**-25.79 ± 5.79**
Claude 2	-	Y	-	-	-	-	-	3.16 ± 6.84
GPT 3.5 T	-	Y	-	4.21 ± 6.84	3.16 ± 6.32	-2.11 ± 11.58	-2.11 ± 11.58	-3.68 ± 8.95
GPT 4	-	Y	-	**12.11 ± 6.84**	3.68 ± 6.84	10.53 ± 10.53	6.32 ± 10.53	1.05 ± 7.37

Table 4: The difference between scores on "I am a Strange Dataset" when the referent is "the following sentence" instead of "this sentence" (scores for the first minus the latter). Overall, the problems that LLMs have with self-referential statements do not stem only from issues understanding the self-referential referent itself. Differences outside of the 95% confidence interval (computed the same way as for Table 2) are shown in **bold**.

mance is well below the non-expert human scores.

Results in this paper for Claude 2 (Anthropic, 2023), GPT 3.5 Turbo (Brown et al., 2020), and GPT 4 (OpenAI, 2023) were collected through the `gpt-4`, `gpt-3.5-turbo`, and `claude-2` endpoints on their respective APIs on Jan 7, 2024 (unless otherwise stated). Note that the Claude 2 API does not support logprobs, so the only metric that we report for it is the text-based CoT validation metric. The OpenAI API supports top-5 logprobs access, and it turns out that this is enough to get logprobs for "`True`" vs "`False`" in our experiments. It is not enough to get the multi-token logprobs required for the generation metric, though.

The CoT metric gives us extra insight into the limitations of models because we can see if their reasoning aligns with their final answer. In many cases, they make fairly obvious mistakes as seen in Figure 3. In some cases, the models choose the correct answer even though their reasoning is flawed, as seen in Figure 5.

It is important to note that there are limitations with the CoT validation metric, and the ZS and FS non-relative validation metrics too. These metrics are about judging a statement as true or false, independent of the statement with the alternative continuation. The dataset schema calls for sentences which are either true or false, but in practice, language can be ambiguous. Figure 6 illustrates. A fairer gauge of model capability may be to compare the continuations with each other, as the generation and relative validation metrics do.

Tags. We present the average logprobs-based model scores for the tags in Table 3. We opt to leave CoT scores out due to their high variance (many of the open-source models do not even return either "`true`" or "`false`" in their answers). Full tables for the tags are in Appendix E.

Non-Self-Referential Control. We provide the difference in scores between the non-self-referent examples and analogous original examples in Table 4. The significant values in the CoT column are not particularly compelling because the open-source models are below chance regardless of whether there is self-reference. For this metric, the open-source models are extremely sensitive to prompting – many of them do not even output "`true`" or "`false`" for every example.

We find that our dataset is so hard due to non-self-referential metalinguistic aspects that models

Why doesn't this question have an answer? It does, though.

GPT 4: The statement is asking why it doesn't have an answer, but it does have an answer. So, the statement is false.

Figure 6: This is a particularly ambiguous example that is assigned the gold label of true. The response from GPT 4 is marked as incorrect, although one could make a case that the statement is false because of the false presupposition which is corrected later in the answer. The generation and relative validation metrics are robust to these situations. For these metrics, the losses from "`It does though.`" versus "`It just doesn't.`" (the alternative continuation) are compared directly. Regardless of ambiguity about whether a statement is true outright, we can notice a higher confidence for "`It just doesn't.`" as the false continuation.

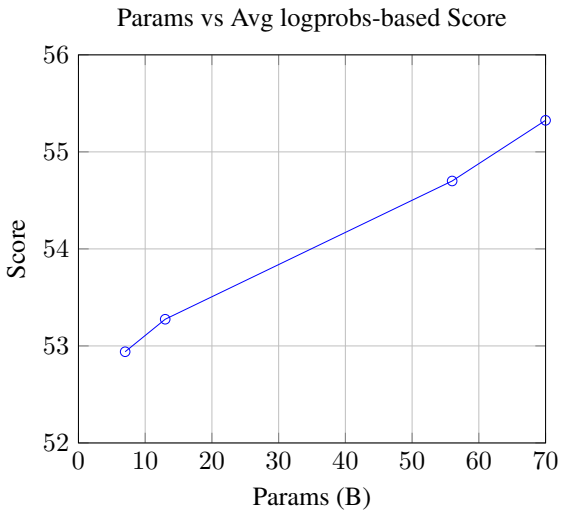

Figure 7: Parameters to average "I am a Strange Dataset" score across all of the logprobs-based metrics. We only evaluate five 7B models, and two models for each of the other sizes, so computing confidence intervals for each point is not particularly informative. Under the null hypothesis that parameter size has no effect on score, we can compute the p-value for these results as a whole nonparametrically: there are 24 ways that these 4 points can be arranged and in only 1 of the ways do they all increase with the parameter count: $p = 1/24 = 0.042$.

score around chance here too, with the exception of GPT 4. It is the only model which is strong enough to perform significantly above chance for every metric tested on the main dataset in Table 2, and to have all positive values in Table 4 for every metric tested, including logprobs-based metrics (meaning that the self-referential version was harder for it). Although, only the Val ZS value for GPT 4 is well outside of the 95% confidence interval, and GPT 4 is also still not particularly good at the non-self-referential version. Figure 4 shows that GPT 4 struggles with the non-self-referent version of the Figure 3 example. There is some signal that the challenge posed by self-referents will remain as LLMs gain competence at other metalinguistic problems, but the dataset is so hard that we do not have overwhelming evidence.

Model Scale. If we exclude the high-variance CoT metric, we see a clear scaling trend that models with more parameters score higher on the test. See Figure 7. Will this trend continue? For additional discussion, see Appendix D.

6 Conclusion

A grasp of self-reference is important in a variety of domains, and is a notable aspect of human intelligence. We introduced a novel task and dataset to assess the capacity of models to generate and understand self-referential statements, and understand the prerequisite metalinguistic reasoning. All models that we tested perform fairly close to the level of chance. GPT 4 is the only model to score significantly above chance on all of the metrics tested, and still it is not by much. The poor performance may be indicative of a larger issue about the limitations of even today's best causal language models. Even though the task is straightforward for people, we find evidence that scale beyond 70B parameters may be needed for the emergence of comparable performance from models.

Our results indicate that this dataset is hard not only due to the self-referent part of a self-referential statement. The challenge also comes from other metalinguistic aspects, such as recursively applying arithmetic operations on sentences. Still, there is some limited evidence that GPT 4 struggles more with self-referential metalinguistic problems than analogous non-self-referential problems.

7 Dataset Release Strategy

We release the dataset on GitHub. The data is encrypted, but the decryption script is provided. Our goal is not to hide the dataset from people, but to hide the dataset from any processes that scrape training data from the web. We encourage the rest of the community to take up this practice when releasing evaluation datasets in a public repository.

8 Limitations

It is possible that the self-reference aspect of "I am a Strange Dataset" will turn out to be the bottleneck for many models, but it is also true that models are largely failing at the purely metalinguistic aspect. Although the schema targets metalinguistic self-reference, it is difficult to make a specific claim about why models fail without running more experiments and without waiting until models become more competent.

9 Ethical Considerations

We aimed to pay crowdworkers 15 USD hourly based on an estimated task completion time.

References

Richard Allan. 2017. Hard questions: Who should decide what is hate speech in an online global community?

Amazon. Retrieved 2023. Mechanical turk.

Anonymous. 2024. I am a strange dataset: Metalinguistic tests for language models. *In Review*.

Anthropic. 2023. Claude.

Bernard J. Baars. 1993. *A cognitive theory of consciousness*. Cambridge University Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv*.

Alonzo Church. 1936. An unsolvable problem of elementary number theory. *American Journal of Mathematics*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020a. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020b. Syntaxgym: An online platform for targeted evaluation of language models. *ACL System Demos*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *NAACL: Human Language Technologies*.

Kurt Gödel. 1931. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme I. *Monatshefte für Mathematik und Physik*.

Douglas Hofstadter. 1979. *Gödel, Escher, Bach: an Eternal Golden Braid*. Basic Books.

Douglas Hofstadter. 1985. *Metamagical Themas: Questing for the Essence of Mind and Pattern*. Basic Books.

Douglas Hofstadter. 2007. *I Am a Strange Loop*. Basic Books.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. *EMNLP*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv*.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Conference on the Principles of Knowledge Representation and Reasoning*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *TACL*.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. 2022. The inverse scaling prize.

OpenAI. 2022. ChatGPT.

OpenAI. 2023. GPT-4 technical report. *arXiv*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *ACL*.

James Propp. 1993. Self-referential aptitude test.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, , and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical Report*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *arXiv preprint arXiv:1804.09301*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*.

SciPy. Retrieved 2023. Scipy bootstrap.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *EMNLP*.

Alfred Tarski. 1931. The concept of truth in formalized languages. *Logic, Semantics, Metamathematics*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. *CVPR*.

Tristan Thrush, Ethan Wilcox, and Roger Levy. 2020. Investigating novel verb learning in BERT: Selectional preference classes and alternation-based syntactic generalization. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. *EMNLP 2023 Findings*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*.

Eric Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. 2023. Self-taught optimizer (stop): Recursively self-improving code generation. *arXiv*.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023. Solving challenging math word problems using GPT-4 code interpreter with code-based self-verification. *arXiv*.

Banghua Zhu, Hiteshi Sharma, Felipe Vieira Frujeri, Shi Dong, Chenguang Zhu, Michael I. Jordan, and Jiantao Jiao. 2023. Fine-tuning language models with advantage-induced policy alignment. *arXiv*.

A Model Prompts

In this section, we provide the full model prompts that we used for the validation metrics. The generation metric does not use a prompt (generation is conditioned on the beginning of a statement). For each prompt, `$example` is replaced by one of the 400 full statements before being given to a model.

A.1 Zero Shot

```
Overview:

Your task is to say whether a self-referential
statement is true or false. In cases where you
think that the answer is ambiguous, you should
use your best judgement and still say either
'True' or 'False'.


Example:

$example

Answer:
```

A.2 Few Shot

```
Overview:

Your task is to say whether a self-referential
statement is true or false. In cases where you
think that the answer is ambiguous, you should
use your best judgement and still say either
'True' or 'False'. Make sure that you understand
the given examples before continuing.


Example:

I am the text you are reading.
My last line has how many words?
Two words.

Answer: True


Example:

I am the text you are reading.
My last line has how many words?
Seven words.

Answer: False


Example:

this sentence
l             o
o             o
k             s
a bit like the
l
e
t
t
e
r
"p"

Answer: True


Example:

this sentence
l             o
o             o
k             s
a bit like the
l
e
t
t
e
r
"o"

Answer: False


Example:

here in reversed is order word the.

Answer: True


Example:

here in fine is order word the.

Answer: False


Now it is your turn to complete an example.
Fill in the answer.

Example:

$example

Answer:
```

A.3 Chain of Thought

```
Overview:

Your task is to say whether a self-referential
statement is true or false. In cases where you
think that the answer is ambiguous, you should
use your best judgement and still say either
'True' or 'False'. Make sure that you understand
the given examples before continuing.


Example:

I am the text you are reading.
My last line has how many words?
Two words.

Answer: The text has 3 lines. The last line is
"Two words". The number of words in "Two words"
is 2. So the statement is true.


Example:
```

```
I am the text you are reading.
My last line has how many words?
Seven words.

Answer: The text has 3 lines. The last line is
"Seven words". The number of words in "Seven words"
is 2. So the statement is false.


Example:

this sentence
l             o
o             o
k             s
a bit like the
l
e
t
t
e
r
"p"

Answer: Visually, the words and letters of the
sentence are arranged to look like a "p" on the
page. So, the statement is true.


Example:

this sentence
l             o
o             o
k             s
a bit like the
l
e
t
t
e
r
"o"

Answer: Visually, the words and letters of the
sentence are arranged to look like a "p" on the
page, not an "o". So, the statement is false.


Example:

here in reversed is order word the.

Answer: We have to notice that, if the word order is
reversed, the statement reads: "the word order is
reversed in here". In the original reversed
statement, that is true. So the statement is true.


Example:

here in fine is order word the.

Answer: We have to notice that, if the word order is
reversed, the statement reads: "the word order is
fine in here". In the original reversed statement,
the word order is not fine. So the statement is
false.
```

```
Now it is your turn to complete an example.
Fill in the answer.

Example:

$example

Answer:
```

A.4 Zero Shot (Non-Self-Referent)

```
Overview:

Your task is to say whether a statement about a
sentence is true or false. In cases where you think
that the answer is ambiguous, you should use your
best judgement and still say either 'True' or
'False'.


Example:

$example

Answer:
```

A.5 Few Shot (Non-Self-Referent)

```
Overview:

Your task is to say whether a statement about a
sentence is true or false. In cases where you think
that the answer is ambiguous, you should use your
best judgement and still say either 'True' or
'False'. Make sure that you understand the given
examples before continuing.


Example:

The number of words in the following sentence is
nine.

The number of words in this sentence is nine.

Answer: True


Example:

The number of words in the following sentence is
eight.

The number of words in this sentence is eight.

Answer: False


Example:

the following sentence
l             o
o             o
k             s
a bit like the
l
e
t
t
e
```

```
r
"p"

this sentence
l             o
o             o
k             s
a bit like the
l
e
t
t
e
r
"p"

Answer: True


Example:

the following sentence
l             o
o             o
k             s
a bit like the
l
e
t
t
e
r
"o"

this sentence
l             o
o             o
k             s
a bit like the
l
e
t
t
e
r
"o"

Answer: False


Example:

The food words that the following sentence contains
are guava and cake and also carrot.

The food words that this sentence contains are
guava and cake and also carrot.

Answer: True


Example:

The food words that the following sentence contains
are guava and cake but not carrot.

The food words that this sentence contains are
guava and cake but not carrot.

Answer: False
```

```
Now it is your turn to complete an example.
Fill in the answer.

Example:

$example

Answer:
```

A.6 Chain of Thought (Non-Self-Referent)

```
Overview:

Your task is to say whether a statement about a
sentence is true or false. In cases where you think
that the answer is ambiguous, you should use your
best judgement and still say either 'True' or
'False'. Make sure that you understand the given
examples before continuing.


Example:

The number of words in the following sentence is
nine.

The number of words in this sentence is nine.

Answer: Counting the number of words in the
sentence, we see that there are nine words. So,
the statement is true.


Example:

The number of words in the following sentence is
eight.

The number of words in this sentence is eight.

Answer: Counting the number of words in the
sentence, we see that there are nine words,
not eight. So, the statement is false.


Example:

the following sentence
l             o
o             o
k             s
a bit like the
l
e
t
t
e
r
"p"

this sentence
l             o
o             o
k             s
a bit like the
l
e
t
```

```
t
e
r
"p"

Answer: Visually, the words and letters of
the sentence are arranged to look like a "p"
on the page. So, the statement is true.


Example:

the following sentence
l             o
o             o
k             s
a bit like the
l
e
t
t
e
r
"o"

this sentence
l             o
o             o
k             s
a bit like the
l
e
t
t
e
r
"o"

Answer: Visually, the words and letters of the
sentence are arranged to look like a "p" on the
page, not an "o". So, the statement is false.


Example:

The food words that the following sentence contains
are guava and cake and also carrot.

The food words that this sentence contains are
guava and cake and also carrot.

Answer: The food words mentioned are indeed guava,
cake, and carrot. So, the statement is true.


Example:

The food words that the following sentence contains
are guava and cake but not carrot.

The food words that this sentence contains are
guava and cake but not carrot.

Answer: The food words mentioned in the sentence
are guava, cake, and carrot. It is not true that
carrot is not in the sentence. So, the statement
is false.


Now it is your turn to complete an example.

Fill in the answer.

Example:

$example

Answer:
```

B I am an Impossible Dataset

The schema of "I am a Strange Dataset" does not place any limits on the complexity of an example. One of the expert annotators in particular took full advantage of this while they were creating examples. They generated 10 examples that another expert annotator could not understand without an explanation. These examples are not in our main dataset, where we verify that non-expert people can perform well. We release these 10 examples separately to showcase just how hard examples fitting this self-referential schema can get. We also hope that when language models improve, these 10 examples can be a useful initial probe into their reasoning capabilities. We provide two examples from this seemingly "Impossible" set in Figure 8. We also provide GPT 4's output on another example from this dataset in Figure 9 below.

Now, I list a digit thirteens: pi. Letter count the words. Conceded? Errorless.	Now, I list a digit thirteens: pi. Letter count the words. Conceded? Erroneous.
Answer: fbdi yqtf lq xlmw hnumjw ynolzy if vul.	Answer: fbdi yqtf lq xlmw hnumjw ynolzy if one.

Figure 8: Examples from "I am an Impossible Dataset". Examples from this dataset still follow our schema, but the expert annotators needed to see an explanation before they understood. The first example says that the letter count of each of the words are the thirteen first digits of π: 3.141592653589. The beginning of the second example uses a Caesar cipher which shifts by one each successive word and reads "each word in this cipher shifts by". "vul" with a shift of 7 is "one".

"Fivetune favors the bold."
"After that meal, I was snined."
"Each of the number sounds in these sentences, including this sentence, are increased by two."

GPT 4: The first sentence "Fivetune favors the bold." has the number sound "five" instead of "fortune". The second sentence "After that meal, I was snined." has the number sound "nine" instead of "sated". The third sentence "Each of the number sounds in these sentences, including this sentence, are increased by two." suggests that each number sound should be increased by two. However, "five" increased by two is "seven", not "fortune", and "nine" increased by two is "eleven", not "sated". So, the statement is false.

Figure 9: GPT 4's Dec 26, 2023 output on an example from "I am an Impossible Dataset". It understands that some of the original words should be "fortune" (e.g. 4tune) and "sated" (e.g. s8ed). But GPT 4 misses that the statement is trying to say that the number sounds in every sentence are increased by one. The last sentence cannot say "one" explicitly - it needs to say "two" in order for the statement to stay true. We would not expect a typical person to understand this example, but will a language model eventually grasp it?

C Mechanical Turk Annotator Interface

Here, we show a screenshot of the interface used by Mechanical Turk workers in Figure 10.

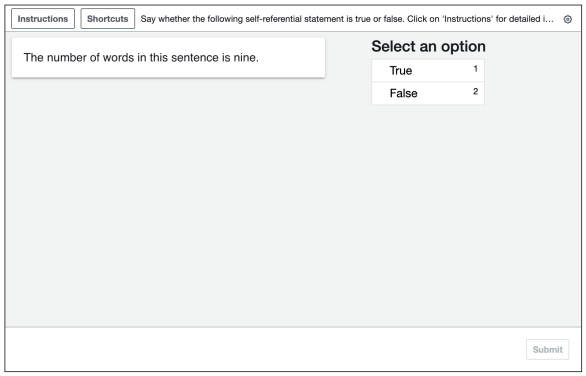

Figure 10: A screenshot of the Mechanical Turk worker interface for validating statements.

D Supplemental Discussion

Here, we present supplemental discussion about why models are showing poor performance on "I am a Strange Dataset".

D.1 A Test of the Tokenizer?

Our tests are related to whether a model truly "sees" text in the same way as people. Metalinguistic statements may refer to the number of characters that they have, how text is arranged on the page, capitalization of certain letters, and relative positions of words. A human can easily count the characters that they see in a sentence, but models tend to encode text in tokens, not characters. We do not provide tests to disentangle the impact of different tokenizers, so this section is speculative.

D.2 Training Data Limitations

Practically speaking, it is unlikely that there are many examples of metalinguistic statements in training datasets. They are incredibly time-intensive to generate, even if they are easy to verify. Yet people, who have almost surely seen even fewer examples, can do much better at this task than models. This "hard-to-create, easy-to-verify" feature of hard evaluation datasets is true of Winoground (Thrush et al., 2022) too, which is a vision and language evaluation dataset that has remained unsaturated for well over a year. This goes to show that our models still have the wrong biases - will they change simply with model scale?

E Results by Tag

In this section, we provide results in a table for each of the 10 tags.

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
Llama 2	7B	N	**62.90 ± 12.10**	49.19 ± 1.21	49.19 ± 4.84	46.77 ± 12.90	50.00 ± 12.90	6.45 ± 4.03
Llama 2	7B	Y	58.06 ± 12.90	50.81 ± 1.21	50.00 ± 0.00	37.10 ± 11.29	56.45 ± 12.90	16.94 ± 6.85
Mistral 0.1	7B	N	53.23 ± 12.90	50.00 ± 2.42	50.81 ± 1.21	56.45 ± 12.90	51.61 ± 12.90	0.00 ± 0.00
Starling α	7B	Y	54.84 ± 12.90	49.19 ± 2.42	49.19 ± 1.21	54.84 ± 12.90	51.61 ± 12.90	32.26 ± 8.87
Mistral 0.2	7B	Y	51.61 ± 12.90	50.00 ± 6.45	50.00 ± 5.65	46.77 ± 12.10	53.23 ± 12.10	53.23 ± 8.06
Llama 2	13B	N	**62.90 ± 11.29**	47.58 ± 3.63	52.42 ± 5.24	41.94 ± 12.10	58.06 ± 12.90	4.03 ± 3.23
Llama 2	13B	Y	59.68 ± 12.90	49.19 ± 4.84	51.61 ± 3.23	46.77 ± 12.90	50.00 ± 12.90	7.26 ± 4.44
Mixtral 0.1	8x7B	N	54.84 ± 12.90	53.23 ± 4.84	50.00 ± 0.00	48.39 ± 12.90	48.39 ± 12.10	3.23 ± 2.82
Mixtral 0.1	8x7B	Y	53.23 ± 12.90	49.19 ± 4.84	45.97 ± 5.65	53.23 ± 12.90	48.39 ± 12.90	36.29 ± 8.06
Llama 2	70B	N	61.29 ± 11.29	52.42 ± 5.65	50.81 ± 4.03	59.68 ± 11.29	50.00 ± 12.90	3.23 ± 2.82
Llama 2	70B	Y	51.61 ± 12.90	53.23 ± 6.45	49.19 ± 3.23	56.45 ± 12.90	46.77 ± 12.90	31.45 ± 7.26
Claude 2	-	Y	-	-	-	-	-	48.39 ± 6.85
GPT 3.5 T	-	Y	-	48.39 ± 4.84	50.81 ± 5.65	56.45 ± 12.90	50.00 ± 12.90	44.35 ± 7.26
GPT 4	-	Y	-	53.23 ± 8.06	53.23 ± 8.06	54.84 ± 12.90	53.23 ± 12.90	55.65 ± 9.27

Table 5: Results for the 62 example pairs with the Numerical Operations tag. Scores with 95% confidence intervals above chance are shown in **bold**.

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
Llama 2	7B	N	54.35 ± 9.78	49.46 ± 2.17	51.63 ± 3.53	47.83 ± 10.33	**66.30 ± 9.78**	3.80 ± 2.72
Llama 2	7B	Y	48.91 ± 9.78	54.35 ± 5.16	50.54 ± 0.82	**60.87 ± 9.78**	**61.96 ± 9.78**	13.04 ± 4.35
Mistral 0.1	7B	N	51.09 ± 9.78	**55.43 ± 4.89**	49.46 ± 2.17	**63.04 ± 9.78**	56.52 ± 9.78	0.00 ± 0.00
Starling α	7B	Y	50.00 ± 9.78	**59.24 ± 5.43**	51.63 ± 2.17	**60.87 ± 9.78**	**61.96 ± 9.78**	38.04 ± 6.79
Mistral 0.2	7B	Y	51.09 ± 9.78	54.35 ± 6.79	52.17 ± 5.98	56.52 ± 9.78	57.61 ± 9.78	50.00 ± 6.52
Llama 2	13B	N	53.26 ± 9.78	54.89 ± 5.98	**56.52 ± 5.98**	55.43 ± 9.78	**65.22 ± 9.78**	5.98 ± 3.26
Llama 2	13B	Y	48.91 ± 9.78	55.98 ± 6.52	52.17 ± 4.35	57.61 ± 9.78	50.00 ± 9.78	11.96 ± 5.16
Mixtral 0.1	8x7B	N	55.43 ± 9.78	**63.59 ± 6.52**	52.72 ± 4.35	**64.13 ± 9.78**	**60.87 ± 9.78**	3.80 ± 2.99
Mixtral 0.1	8x7B	Y	55.43 ± 10.33	**57.07 ± 6.52**	**57.07 ± 4.62**	**61.96 ± 9.78**	57.61 ± 10.33	50.00 ± 7.07
Llama 2	70B	N	55.43 ± 9.78	**57.07 ± 5.43**	**59.24 ± 4.89**	**66.30 ± 9.78**	**63.04 ± 9.78**	1.09 ± 1.36
Llama 2	70B	Y	51.09 ± 10.33	**59.24 ± 7.07**	52.72 ± 3.80	**61.96 ± 9.78**	**65.22 ± 9.78**	21.74 ± 6.25
Claude 2	-	Y	-	-	-	-	-	55.98 ± 7.07
GPT 3.5 T	-	Y	-	54.89 ± 4.89	**55.98 ± 5.72**	57.61 ± 9.78	**71.74 ± 9.24**	55.98 ± 6.25
GPT 4	-	Y	-	**63.04 ± 6.52**	**61.96 ± 6.52**	**70.65 ± 9.24**	**71.74 ± 9.24**	**72.83 ± 6.52**

Table 6: Results for the 94 example pairs with the Negation & Scope tag. Scores with 95% confidence intervals above chance are shown in **bold**.

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
Llama 2	7B	N	45.83 ± 20.83	50.00 ± 0.00	50.00 ± 6.25	58.33 ± 20.83	58.33 ± 20.83	8.33 ± 7.29
Llama 2	7B	Y	45.83 ± 20.83	52.08 ± 6.25	50.00 ± 0.00	62.50 ± 20.83	**70.83 ± 18.75**	22.92 ± 10.42
Mistral 0.1	7B	N	45.83 ± 20.83	52.08 ± 6.25	54.17 ± 5.21	66.67 ± 18.75	**70.83 ± 18.75**	0.00 ± 0.00
Starling α	7B	Y	45.83 ± 20.83	56.25 ± 10.42	47.92 ± 3.12	66.67 ± 18.75	66.67 ± 18.75	37.50 ± 12.50
Mistral 0.2	7B	Y	54.17 ± 20.83	54.17 ± 10.42	62.50 ± 12.50	62.50 ± 20.83	66.67 ± 18.75	47.92 ± 12.50
Llama 2	13B	N	45.83 ± 20.83	52.08 ± 6.25	56.25 ± 8.33	54.17 ± 20.83	58.33 ± 20.83	2.08 ± 3.12
Llama 2	13B	Y	45.83 ± 20.83	52.08 ± 10.42	52.08 ± 3.12	58.33 ± 20.83	66.67 ± 18.75	18.75 ± 11.46
Mixtral 0.1	8x7B	N	37.50 ± 18.75	**64.58 ± 10.42**	54.17 ± 5.21	62.50 ± 18.75	62.50 ± 20.83	12.50 ± 9.38
Mixtral 0.1	8x7B	Y	50.00 ± 20.83	47.92 ± 8.33	54.17 ± 10.42	50.00 ± 20.83	58.33 ± 20.83	41.67 ± 14.58
Llama 2	70B	N	45.83 ± 20.83	50.00 ± 8.33	56.25 ± 12.50	54.17 ± 20.83	62.50 ± 18.75	12.50 ± 8.33
Llama 2	70B	Y	50.00 ± 20.83	43.75 ± 10.42	47.92 ± 3.12	54.17 ± 20.83	58.33 ± 20.83	10.42 ± 8.33
Claude 2	-	Y	-	-	-	-	-	47.92 ± 12.50
GPT 3.5 T	-	Y	-	50.00 ± 0.00	47.92 ± 6.25	54.17 ± 20.83	58.33 ± 20.83	52.08 ± 12.50
GPT 4	-	Y	-	54.17 ± 10.42	**66.67 ± 13.54**	58.33 ± 20.83	62.50 ± 18.75	60.42 ± 14.58

Table 7: Results for the 24 example pairs with the Multi-Channel tag. Scores with 95% confidence intervals above chance are shown in **bold**.

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
Llama 2	7B	N	50.00 ± 12.96	48.15 ± 2.31	48.15 ± 3.70	35.19 ± 12.96	46.30 ± 12.96	6.48 ± 4.17
Llama 2	7B	Y	44.44 ± 12.96	48.15 ± 4.63	49.07 ± 1.39	46.30 ± 12.96	38.89 ± 12.96	12.04 ± 5.56
Mistral 0.1	7B	N	46.30 ± 12.96	47.22 ± 3.24	47.22 ± 3.24	37.04 ± 12.96	44.44 ± 12.96	0.00 ± 0.00
Starling α	7B	Y	48.15 ± 12.96	48.15 ± 5.09	50.93 ± 1.39	44.44 ± 12.96	46.30 ± 12.96	28.70 ± 8.33
Mistral 0.2	7B	Y	46.30 ± 12.96	52.78 ± 7.41	48.15 ± 7.41	44.44 ± 12.96	37.04 ± 12.96	43.52 ± 8.33
Llama 2	13B	N	48.15 ± 12.96	48.15 ± 5.56	49.07 ± 5.56	44.44 ± 12.96	44.44 ± 12.96	3.70 ± 3.24
Llama 2	13B	Y	44.44 ± 12.96	47.22 ± 7.41	48.15 ± 4.63	38.89 ± 12.96	35.19 ± 12.96	8.33 ± 5.09
Mixtral 0.1	8x7B	N	46.30 ± 12.96	54.63 ± 5.56	49.07 ± 3.70	44.44 ± 12.96	46.30 ± 12.96	1.85 ± 2.31
Mixtral 0.1	8x7B	Y	44.44 ± 12.96	51.85 ± 8.33	50.00 ± 6.48	44.44 ± 12.96	53.70 ± 12.96	44.44 ± 9.26
Llama 2	70B	N	50.00 ± 12.96	50.00 ± 4.63	50.93 ± 3.70	51.85 ± 12.96	44.44 ± 12.96	0.00 ± 0.00
Llama 2	70B	Y	44.44 ± 12.96	46.30 ± 7.41	47.22 ± 4.17	44.44 ± 12.96	51.85 ± 12.96	29.63 ± 8.33
Claude 2	-	Y	-	-	-	-	-	50.00 ± 7.41
GPT 3.5 T	-	Y	-	51.85 ± 5.09	46.30 ± 6.48	51.85 ± 12.96	55.56 ± 12.96	54.63 ± 9.26
GPT 4	-	Y	-	56.48 ± 7.41	58.33 ± 9.26	**64.81 ± 12.96**	62.96 ± 12.96	**71.30 ± 8.80**

Table 8: Results for the 55 example pairs with the Location of Element tag. Scores with 95% confidence intervals above chance are shown in **bold**.

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
Llama 2	7B	N	**66.67 ± 14.29**	50.00 ± 0.00	47.62 ± 4.76	52.38 ± 14.29	52.38 ± 14.29	7.14 ± 5.36
Llama 2	7B	Y	61.90 ± 14.29	54.76 ± 5.36	48.81 ± 1.79	54.76 ± 14.29	52.38 ± 14.29	25.00 ± 8.93
Mistral 0.1	7B	N	**66.67 ± 14.29**	50.00 ± 0.00	47.62 ± 2.98	52.38 ± 14.29	52.38 ± 14.29	0.00 ± 0.00
Starling α	7B	Y	**66.67 ± 14.29**	53.57 ± 5.36	51.19 ± 3.57	59.52 ± 14.29	54.76 ± 14.29	32.14 ± 9.52
Mistral 0.2	7B	Y	57.14 ± 14.29	48.81 ± 8.33	54.76 ± 7.14	50.00 ± 14.29	50.00 ± 14.29	55.95 ± 8.33
Llama 2	13B	N	**73.81 ± 13.10**	45.24 ± 5.95	54.76 ± 5.36	52.38 ± 14.29	54.76 ± 14.29	3.57 ± 4.17
Llama 2	13B	Y	**71.43 ± 13.10**	55.95 ± 8.33	48.81 ± 1.79	59.52 ± 14.29	52.38 ± 14.29	13.10 ± 6.55
Mixtral 0.1	8x7B	N	57.14 ± 14.29	55.95 ± 7.14	46.43 ± 4.17	52.38 ± 14.29	54.76 ± 14.29	3.57 ± 4.17
Mixtral 0.1	8x7B	Y	57.14 ± 14.29	53.57 ± 7.14	52.38 ± 6.55	42.86 ± 14.29	54.76 ± 14.29	34.52 ± 8.33
Llama 2	70B	N	**69.05 ± 14.29**	52.38 ± 7.14	**61.90 ± 6.55**	61.90 ± 14.29	61.90 ± 14.29	2.38 ± 2.98
Llama 2	70B	Y	**64.29 ± 14.29**	55.95 ± 8.33	50.00 ± 3.57	59.52 ± 14.29	61.90 ± 14.29	13.10 ± 7.74
Claude 2	-	Y	-	-	-	-	-	**59.52 ± 8.33**
GPT 3.5 T	-	Y	-	48.81 ± 4.76	53.57 ± 7.14	**64.29 ± 14.29**	61.90 ± 14.29	47.62 ± 9.52
GPT 4	-	Y	-	50.00 ± 7.14	**58.33 ± 7.14**	59.52 ± 14.29	59.52 ± 14.29	55.95 ± 9.52

Table 9: Results for the 42 example pairs with the Sensory tag. Scores with 95% confidence intervals above chance are shown in **bold**.

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
Llama 2	7B	N	62.50 ± 13.54	50.00 ± 0.00	48.96 ± 6.25	41.67 ± 14.58	41.67 ± 14.58	9.38 ± 5.21
Llama 2	7B	Y	58.33 ± 14.58	48.96 ± 6.25	50.00 ± 0.00	31.25 ± 12.50	62.50 ± 14.58	8.33 ± 5.21
Mistral 0.1	7B	N	62.50 ± 13.54	46.88 ± 4.17	48.96 ± 3.12	41.67 ± 13.54	43.75 ± 14.58	0.00 ± 0.00
Starling α	7B	Y	**64.58 ± 13.54**	48.96 ± 3.12	48.96 ± 1.56	39.58 ± 13.54	52.08 ± 14.58	31.25 ± 8.85
Mistral 0.2	7B	Y	**64.58 ± 13.54**	45.83 ± 8.33	50.00 ± 6.25	47.92 ± 14.58	41.67 ± 14.58	46.88 ± 8.33
Llama 2	13B	N	60.42 ± 13.54	48.96 ± 6.25	47.92 ± 8.33	43.75 ± 14.58	47.92 ± 14.58	5.21 ± 4.17
Llama 2	13B	Y	56.25 ± 14.58	46.88 ± 7.29	48.96 ± 4.69	45.83 ± 14.58	39.58 ± 14.58	5.21 ± 4.17
Mixtral 0.1	8x7B	N	54.17 ± 14.58	52.08 ± 6.25	50.00 ± 0.00	33.33 ± 13.54	52.08 ± 14.58	2.08 ± 2.60
Mixtral 0.1	8x7B	Y	60.42 ± 14.58	52.08 ± 7.29	52.08 ± 6.25	50.00 ± 14.58	60.42 ± 14.58	46.88 ± 8.33
Llama 2	70B	N	60.42 ± 13.54	47.92 ± 6.25	53.12 ± 5.21	52.08 ± 14.58	47.92 ± 14.58	1.04 ± 1.56
Llama 2	70B	Y	54.17 ± 14.58	50.00 ± 7.29	47.92 ± 4.17	41.67 ± 14.58	56.25 ± 13.54	23.96 ± 8.33
Claude 2	-	Y	-	-	-	-	-	57.29 ± 7.29
GPT 3.5 T	-	Y	-	52.08 ± 4.17	53.12 ± 6.25	54.17 ± 14.58	60.42 ± 13.54	45.83 ± 9.38
GPT 4	-	Y	-	**61.46 ± 6.77**	**62.50 ± 7.81**	**72.92 ± 12.50**	**70.83 ± 12.50**	**62.50 ± 9.38**

Table 10: Results for the 48 example pairs with the Sub-Word tag. Scores with 95% confidence intervals above chance are shown in **bold**.

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
Llama 2	7B	N	64.52 ± 16.13	51.61 ± 2.42	53.23 ± 6.45	54.84 ± 16.13	**70.97 ± 16.13**	4.84 ± 5.65
Llama 2	7B	Y	54.84 ± 16.13	56.45 ± 7.26	50.00 ± 0.00	**67.74 ± 16.13**	54.84 ± 16.13	11.29 ± 7.26
Mistral 0.1	7B	N	58.06 ± 16.13	**58.06 ± 6.45**	51.61 ± 2.42	**67.74 ± 16.13**	**67.74 ± 16.13**	0.00 ± 0.00
Starling α	7B	Y	51.61 ± 16.13	**62.90 ± 7.26**	53.23 ± 4.03	**67.74 ± 16.13**	54.84 ± 17.74	46.77 ± 11.29
Mistral 0.2	7B	Y	61.29 ± 16.13	53.23 ± 12.90	54.84 ± 9.68	64.52 ± 16.13	61.29 ± 16.13	53.23 ± 11.29
Llama 2	13B	N	**67.74 ± 16.13**	59.68 ± 9.68	59.68 ± 10.48	**67.74 ± 16.13**	**77.42 ± 14.52**	8.06 ± 6.45
Llama 2	13B	Y	58.06 ± 16.13	**62.90 ± 9.68**	**56.45 ± 5.65**	**67.74 ± 16.13**	64.52 ± 16.13	8.06 ± 7.26
Mixtral 0.1	8x7B	N	58.06 ± 16.13	**69.35 ± 9.68**	54.84 ± 4.84	**74.19 ± 16.13**	61.29 ± 16.13	1.61 ± 2.42
Mixtral 0.1	8x7B	Y	54.84 ± 17.74	50.00 ± 9.68	54.84 ± 6.45	58.06 ± 16.13	38.71 ± 16.13	41.94 ± 9.68
Llama 2	70B	N	64.52 ± 16.13	**66.13 ± 9.68**	**62.90 ± 7.26**	**67.74 ± 16.13**	**67.74 ± 16.13**	1.61 ± 2.42
Llama 2	70B	Y	61.29 ± 17.74	**67.74 ± 12.10**	50.00 ± 4.84	**70.97 ± 16.13**	**74.19 ± 14.52**	25.81 ± 11.29
Claude 2	-	Y	-	-	-	-	-	59.68 ± 9.68
GPT 3.5 T	-	Y	-	54.84 ± 5.65	**64.52 ± 8.06**	58.06 ± 16.13	64.52 ± 16.13	46.77 ± 12.10
GPT 4	-	Y	-	**64.52 ± 11.29**	61.29 ± 11.29	**67.74 ± 16.13**	**70.97 ± 16.13**	**75.81 ± 12.10**

Table 11: Results for the 31 example pairs with the Existence of Element tag. Scores with 95% confidence intervals above chance are shown in **bold**.

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
Llama 2	7B	N	58.33 ± 20.83	50.00 ± 0.00	50.00 ± 0.00	41.67 ± 20.83	41.67 ± 20.83	6.25 ± 6.25
Llama 2	7B	Y	58.33 ± 18.75	43.75 ± 6.25	50.00 ± 0.00	54.17 ± 20.83	58.33 ± 20.83	29.17 ± 10.42
Mistral 0.1	7B	N	41.67 ± 20.83	43.75 ± 7.29	52.08 ± 3.12	45.83 ± 20.83	54.17 ± 20.83	0.00 ± 0.00
Starling α	7B	Y	37.50 ± 20.83	50.00 ± 6.25	47.92 ± 3.12	45.83 ± 20.83	58.33 ± 20.83	22.92 ± 12.50
Mistral 0.2	7B	Y	45.83 ± 20.83	50.00 ± 10.42	47.92 ± 8.33	41.67 ± 20.83	50.00 ± 20.83	41.67 ± 12.50
Llama 2	13B	N	58.33 ± 20.83	45.83 ± 5.21	54.17 ± 8.33	25.00 ± 16.67	50.00 ± 20.83	4.17 ± 5.21
Llama 2	13B	Y	54.17 ± 20.83	50.00 ± 8.33	47.92 ± 3.12	50.00 ± 20.83	41.67 ± 20.83	16.67 ± 9.38
Mixtral 0.1	8x7B	N	50.00 ± 20.83	50.00 ± 8.33	47.92 ± 3.12	45.83 ± 20.83	50.00 ± 20.83	2.08 ± 3.12
Mixtral 0.1	8x7B	Y	41.67 ± 20.83	43.75 ± 8.33	47.92 ± 10.42	37.50 ± 18.75	37.50 ± 18.85	29.17 ± 12.50
Llama 2	70B	N	50.00 ± 20.83	45.83 ± 8.33	54.17 ± 10.42	41.67 ± 20.83	45.83 ± 20.83	4.17 ± 5.21
Llama 2	70B	Y	54.17 ± 20.83	52.08 ± 7.29	45.83 ± 5.21	41.67 ± 20.83	54.17 ± 20.83	22.92 ± 10.42
Claude 2	-	Y	-	-	-	-	-	41.67 ± 10.42
GPT 3.5 T	-	Y	-	54.17 ± 5.21	45.83 ± 8.33	**70.83 ± 16.67**	54.17 ± 20.83	50.00 ± 10.42
GPT 4	-	Y	-	47.92 ± 9.38	52.08 ± 14.58	41.67 ± 20.83	62.50 ± 18.75	45.83 ± 12.50

Table 12: Results for the 24 example pairs with the Hypothetical tag. Scores with 95% confidence intervals above chance are shown in **bold**.

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
Llama 2	7B	N	44.00 ± 20.00	50.00 ± 6.00	50.00 ± 6.00	40.00 ± 20.00	56.00 ± 20.00	2.00 ± 3.00
Llama 2	7B	Y	60.00 ± 20.00	50.00 ± 10.00	50.00 ± 0.00	52.00 ± 20.00	56.00 ± 20.00	6.00 ± 6.00
Mistral 0.1	7B	N	64.00 ± 18.00	50.00 ± 8.00	46.00 ± 5.00	52.00 ± 20.00	60.00 ± 20.00	0.00 ± 0.00
Starling α	7B	Y	64.00 ± 20.00	54.00 ± 10.00	50.00 ± 0.00	56.00 ± 20.00	60.00 ± 20.00	44.00 ± 15.00
Mistral 0.2	7B	Y	48.00 ± 20.00	56.00 ± 12.00	58.00 ± 10.00	60.00 ± 20.00	60.00 ± 20.00	58.00 ± 12.00
Llama 2	13B	N	60.00 ± 20.00	56.00 ± 8.00	50.00 ± 12.00	52.00 ± 20.00	**68.00 ± 18.00**	6.00 ± 7.00
Llama 2	13B	Y	56.00 ± 20.00	54.00 ± 12.00	54.00 ± 8.00	52.00 ± 20.00	52.00 ± 20.00	8.00 ± 9.00
Mixtral 0.1	8x7B	N	60.00 ± 20.00	58.00 ± 10.00	54.00 ± 8.00	64.00 ± 20.00	52.00 ± 20.00	2.00 ± 3.00
Mixtral 0.1	8x7B	Y	60.00 ± 20.00	44.00 ± 12.00	54.00 ± 5.00	60.00 ± 20.00	56.00 ± 20.00	52.00 ± 16.00
Llama 2	70B	N	56.00 ± 20.00	**60.00 ± 8.00**	56.00 ± 6.00	56.00 ± 20.00	56.00 ± 20.00	0.00 ± 0.00
Llama 2	70B	Y	56.00 ± 20.00	52.00 ± 10.00	48.00 ± 3.00	56.00 ± 20.00	56.00 ± 20.00	30.00 ± 10.00
Claude 2	-	Y	-	-	-	-	-	**68.00 ± 10.00**
GPT 3.5 T	-	Y	-	56.00 ± 8.00	54.00 ± 8.00	56.00 ± 20.00	**72.00 ± 18.00**	42.00 ± 13.05
GPT 4	-	Y	-	**68.00 ± 13.00**	**64.00 ± 13.00**	**84.00 ± 14.00**	**80.00 ± 16.00**	**80.00 ± 11.00**

Table 13: Results for the 25 example pairs with the Grammaticality tag. Scores with 95% confidence intervals above chance are shown in **bold**.

Model	Params	Chat	Gen^L	Val ZS^L	Val FS^L	Val ZS^L (R)	Val FS^L (R)	Val CoT^T
Llama 2	7B	N	59.09 ± 20.45	52.27 ± 6.82	59.09 ± 10.23	59.09 ± 18.18	54.55 ± 20.45	2.27 ± 3.41
Llama 2	7B	Y	59.09 ± 20.45	63.64 ± 13.64	52.27 ± 3.41	59.09 ± 18.18	54.55 ± 22.73	20.45 ± 10.23
Mistral 0.1	7B	N	36.36 ± 20.45	**63.64 ± 11.36**	50.00 ± 0.00	**86.36 ± 13.64**	59.09 ± 20.45	0.00 ± 0.00
Starling α	7B	Y	50.00 ± 20.45	**63.64 ± 11.36**	52.27 ± 6.82	63.64 ± 20.45	54.55 ± 22.73	38.64 ± 13.64
Mistral 0.2	7B	Y	54.55 ± 22.73	59.09 ± 14.77	50.00 ± 11.36	63.64 ± 20.45	68.18 ± 18.18	43.18 ± 12.50
Llama 2	13B	N	59.09 ± 20.45	61.36 ± 13.64	**70.45 ± 13.64**	68.18 ± 18.18	**77.27 ± 18.18**	2.27 ± 3.41
Llama 2	13B	Y	63.64 ± 20.45	56.82 ± 13.64	59.09 ± 11.36	59.09 ± 20.45	63.64 ± 18.18	13.64 ± 12.50
Mixtral 0.1	8x7B	N	59.09 ± 20.45	**72.73 ± 14.77**	**65.91 ± 9.09**	**72.73 ± 18.18**	68.18 ± 18.18	0.00 ± 0.00
Mixtral 0.1	8x7B	Y	45.45 ± 20.45	61.36 ± 15.91	**65.91 ± 9.09**	**72.73 ± 18.18**	68.18 ± 18.18	52.27 ± 13.64
Llama 2	70B	N	63.64 ± 20.45	59.09 ± 13.64	54.55 ± 11.36	**72.73 ± 18.18**	**72.73 ± 18.18**	0.00 ± 0.00
Llama 2	70B	Y	63.64 ± 20.45	**68.18 ± 14.77**	56.82 ± 9.09	**72.73 ± 18.18**	68.18 ± 18.18	9.09 ± 10.23
Claude 2	-	Y	-	-	-	-	-	47.73 ± 13.64
GPT 3.5 T	-	Y	-	**65.91 ± 13.64**	63.64 ± 13.64	59.09 ± 20.45	**72.73 ± 18.18**	**63.64 ± 11.36**
GPT 4	-	Y	-	61.36 ± 11.36	59.09 ± 13.64	**86.36 ± 13.64**	**81.82 ± 15.91**	**72.73 ± 10.23**

Table 14: Results for the 22 example pairs with the Question tag. Scores with 95% confidence intervals above chance are shown in **bold**.