Is ChatGPT-2 Sensitive to Accurate Text Prediction Based on Logic?

Christian DerManuelian
cdermanuelian@ucsd.edu

March 21, 2024

Abstract

AI is really good at what it does. Image recognizers label images with extreme accuracy. State-of-the-art prediction models have given businesses new insights that they could not have dreamed of acquiring before. Chatbots are such good general purpose problem solvers that people pay monthly fees to be able to use them. But why are they good? Do they "think" like humans do, or are they just so mathematically complex and developed that they mimic our logical abilities? We will test this idea on ChatGPT-2. Is this LLM able to make word suggestions and predictions to complete sentences based only on logical connections. Will the chatbot give us the correct answer that only could only have been deduced through logically-implied connections? This will help us think about whether these useful models have any facets of reasoning.

1 Introduction

At its core, any version of ChatGPT is a text generator. Despite its complexity, its principal function is to predict words based on previous prompts and its training data. The question is whether a system like this is capable of any form of logical reasoning. If the answer to that questions is yes, then we would assume that its word selection ability should be modified by logical context. Will language models handle text generation in similar cases differently based on some given context?

2 Methods

2.1 Experimental Design

To understand the experimental design we will use the below variables, used only to illustrate the design.

- S - Some situation
- A - A likely follow up word to S
- B - An unlikely follow up word to S
- U - Context logically implying that B is correct answer, and that A is wrong

A key feature of this design is that A becomes incorrect for the situation when the context U is introduced. With the above variables, the stimuli will follow the below form where we have two prompts.

- Prompt 1: S (A, B)
- Prompt 2: U. S (A, B)

To make this more clear, see the example below.

- Prompt 1: Amelia went to the burger shop to buy a (burger, pizza).
- Prompt 2: Amelia only likes the burger shop's pizza. Amelia went to the burger shop to buy a (burger, pizza).

Without the context, the obvious answer is A, or "burger." Clearly, someone going to a burger shop is there to buy a burger. But any human presented with the fact that Amelia only likes the shop's pizza would then change their answer to B, or "pizza." The motivation behind this is to see if language models are sensitive to this logic when determining the follow-up word to situation S.
All of the prompts will follow this logical style, where the first follow up word is something that the language model would predict without context U. The latter follow up word(s) are unlikely to be predicted without context U. The introduction of this context should switch the model's preference between the two sets of follow-up words.

In the above example, the model would default to the word "burger" because its data about burger stores has a lot of text about people buying burgers, as this is the standard thing to do at a burger store. Then we add the fact that our subject (Amelia) only likes to buy pizza from the store to see if the model catches this connection.

Therefore, a good prompt is one where A has very low surprisal compared to B without context. Surprisal is the negative log of a probability. A high suprisal corresponds to a low probability of the model choosing a certain word. So without context, m surprisal(A) $<$ surprisal(B). Or, "A is a less surprising follow-up word than B."

To test if the model is sensitive to logic, we want to see if the introduction of context U inverts the surprisals of A and B, i.e. surprisal(A) $>$ surprisal(B). Or, "with this context, the previously unsurprising word A is now more surprising than the previously more surprising word B."

Adding surprisal to this makes the explanation a lot more confusing, but it is a requirement of the assignment.

2.2 Experiment 1

Experiment 1 uses the prompt style above, but with two categories; easy and medium. The categories only change the context part of the prompts. Easy prompts directly use the follow up words. The medium prompts still obviously imply the follow up word B, but without using the follow up words and only having logical connections. Again, see the example below.

- Prompt 1: Amelia went to the burger shop to buy a (burger, pizza).
- (Easy) Prompt 2: Amelia only likes the burger shop's pizza. Amelia went to the burger shop to buy a (burger, pizza).
- (Medium) Prompt 2: Amelia only likes the burger shop's Italian food. Amelia went to the burger shop to buy a (burger, pizza).

In experiment 1, we have 6 easy prompts and 6 medium prompts. Each of the 6 prompts between the categories are the same scenario but with different contexts. The aim of having different categories is to see how much help the language model needs to make the logical conclusion. In both categories, we will test the surprisal of the two follow up words once without context, and again with context. If the model shows logical skills, the surprisal should

invert. More specifically, without context surprisal(A) < surprisal(B). With context, surprisal(A) > surprisal(B). Surprisal of course being how unlikely the word is.

2.3 Experiment 2

Experiment 2 is hard prompts. There isn't a quantifiable reason for a prompt being hard. Rather, it is simply a deeper and more convoluded logical connetion that the observer of the prompts must make to pick the correct follow-up word. Additionally, we now have two unlikely words.

- Prompt 1: Mark pulled out a nail using a (hammer, fork, spoon)
- Prompt 2: Mark is in the kitchen because he lost his toolbox. Mark pulled out a nail using a (hammer, fork, spoon)

2.4 Prompts

A detailed list of prompts can be found here.

3 Results

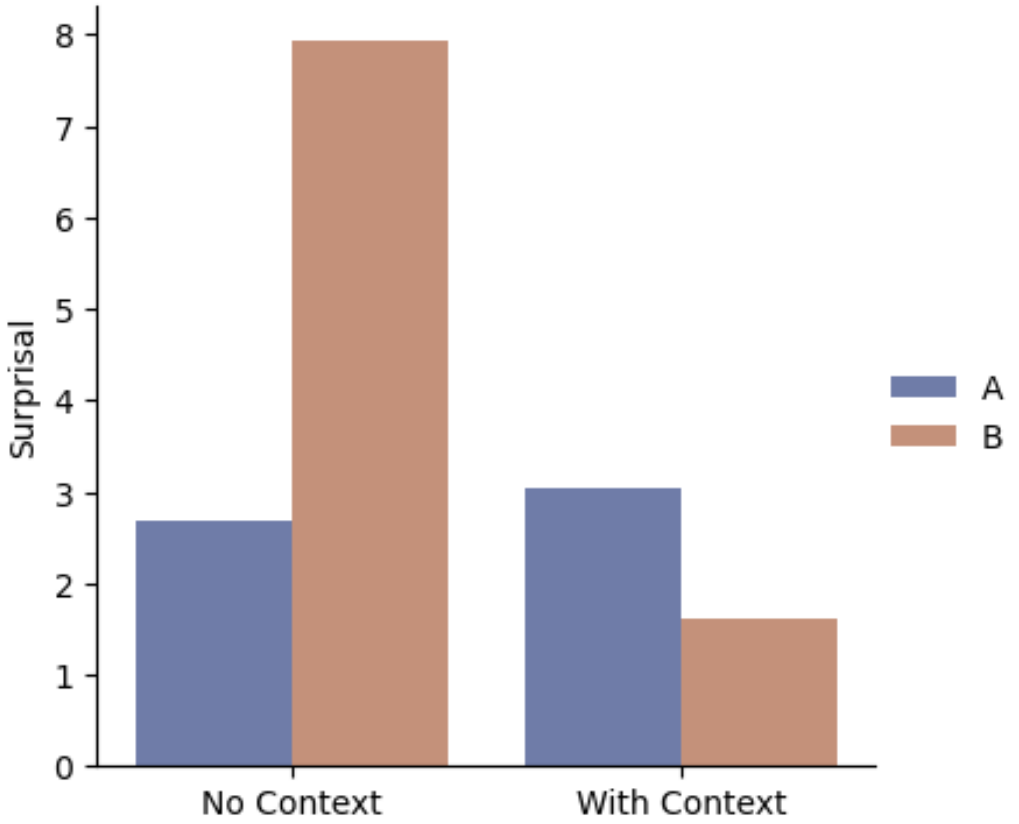
The experiments use the prompts with and without context with ChatGPT-2 to obtain surprisal values for each word (A, B). The code to obtain the suprisal data can be found here. The code to interpet the data and create visualizations can be found here.

3.1 Experiment 1

Recall the example,

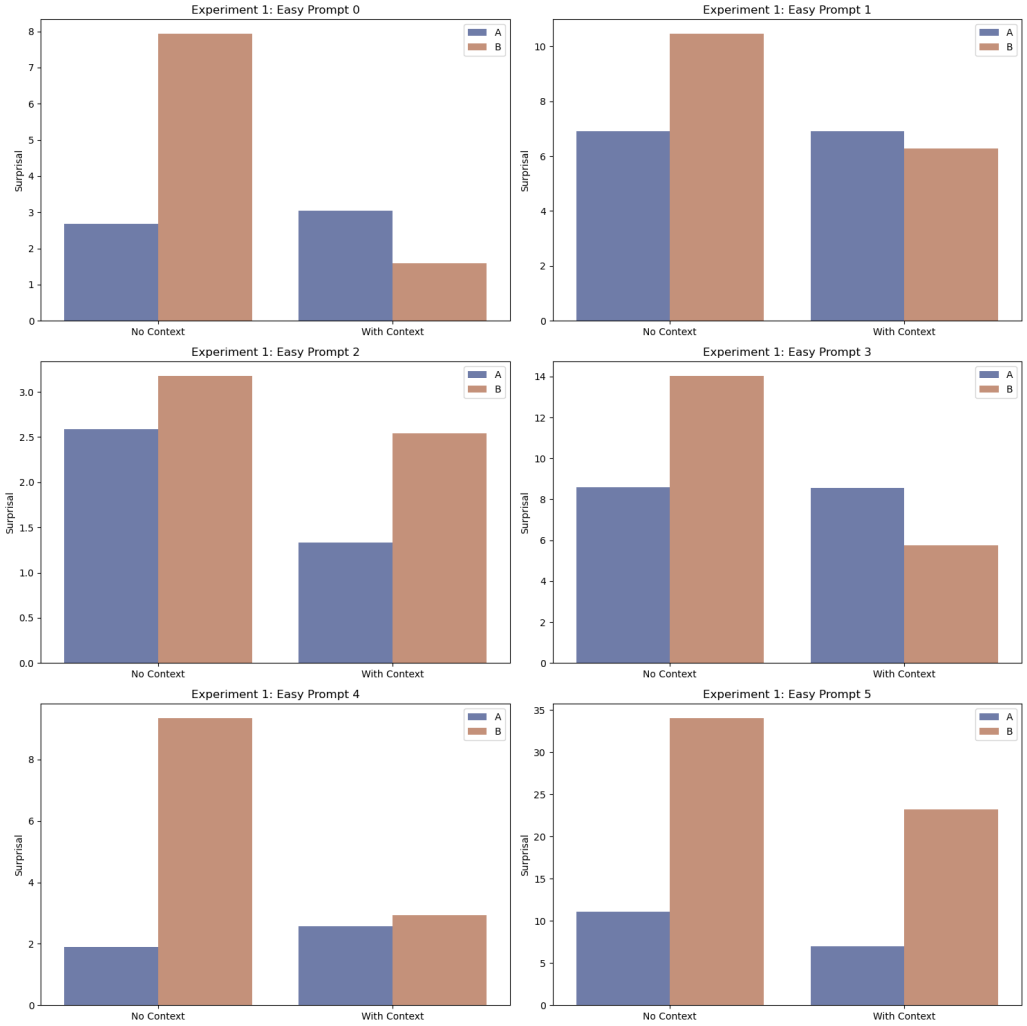
- Prompt 1: Amelia went to the burger shop to buy a (burger, pizza).
- Prompt 2: Amelia only likes the burger shop's pizza. Amelia went to the burger shop to buy a (burger, pizza).

In an ideal case, the surprisal scores for "burger" and "pizza" should switch when context is introduced. In this example, the model actually exhibits the correct interpretation, in which the visualization looks like below.



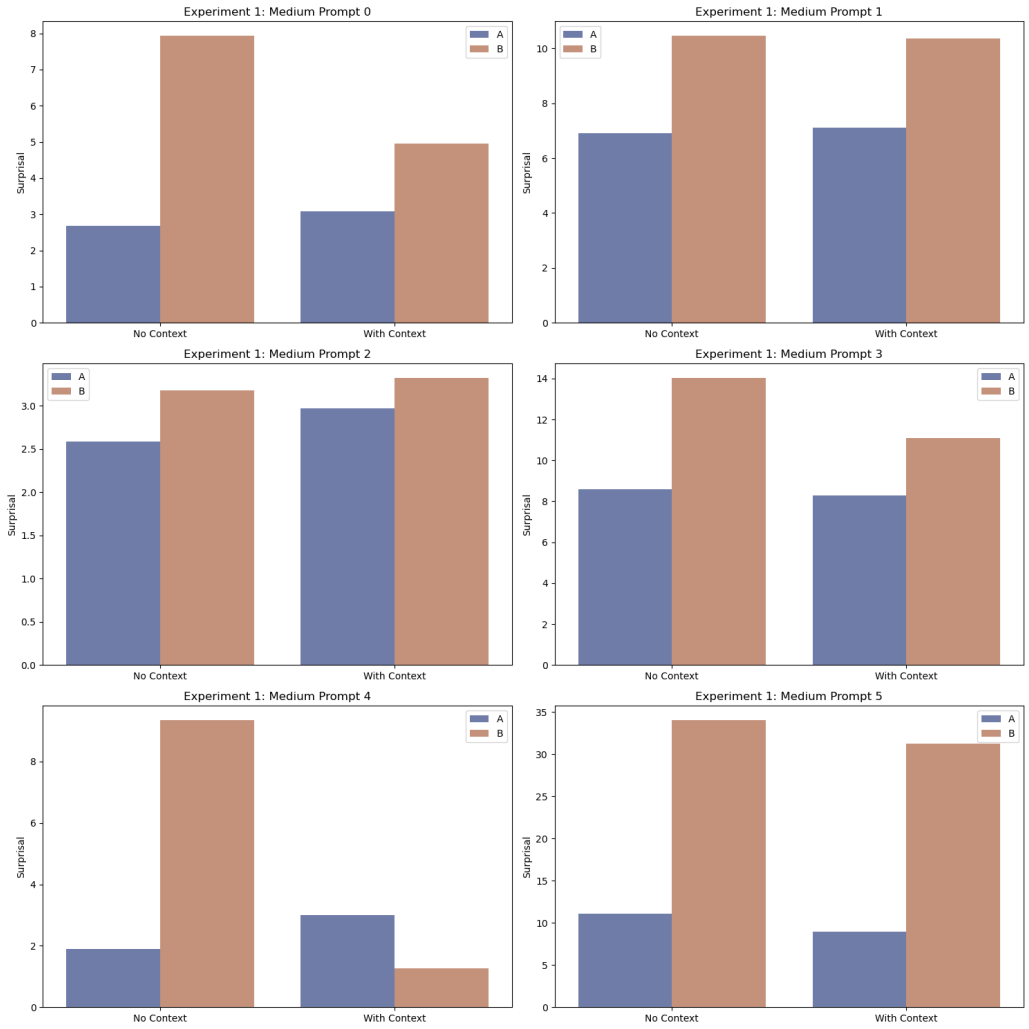
The surprisal for word A (burger) starts out very low relative to word B (pizza). This makes sense as without context, word A is a lot more likely. However, when we introduce context implying B and not A, the surprisal for word A increases, while the surprisal for word B drastically decreases. This means with context, the model is much more likely to finish the sentence with the correct word B.

Now let's see the results for all of the prompts. We will start by analyzing the easy prompts of experiment 1.



Naturally, when there is no context, the bar for surprisal of word A is lower than the surprisal for the bar for word B in all of the plots. What we are interested in is if this inverts when context is introduced. Unfortunately, we see this inversion in only 3 of the prompts. This means that for half of the prompts, ChatGPT-2 is unable to make the correct conclusion with the logical context given. Recall that these are easy prompts, all of which are effortlessly solvable by humans.

What happens when we increase the prompt difficulty?



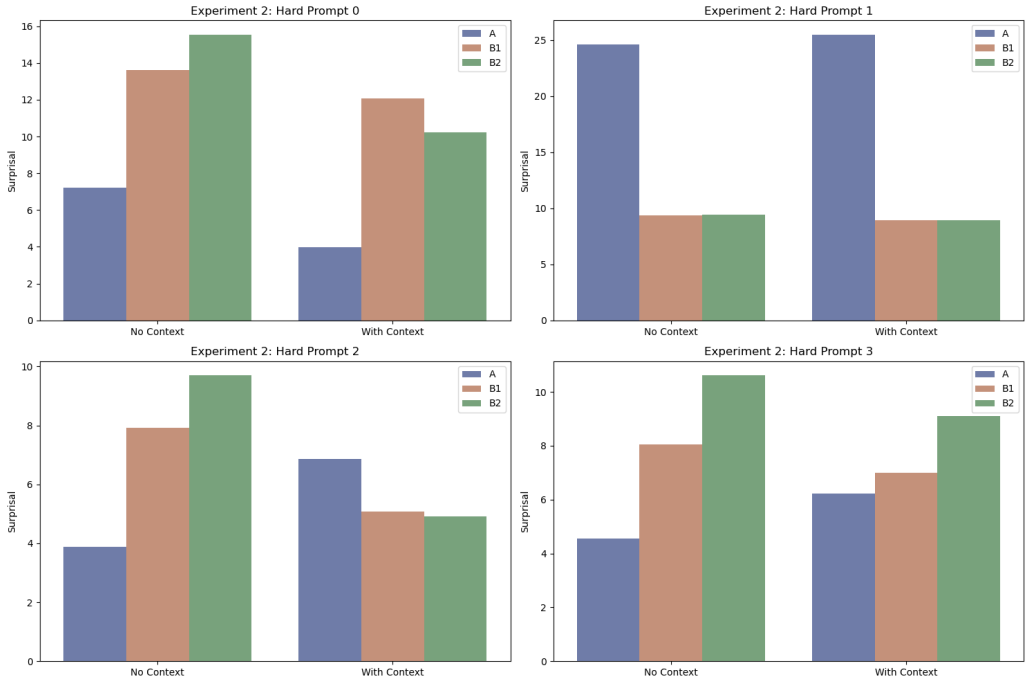
In this case it is even worse, as only one of the prompts was interpreted correctly. It seems as though when the context does not include the actual word, the model shows no ability to make the logical connection.

Interestingly, take a look at prompt 4, the only one that the model got right. The drop in word B's surprisal was very intense. This was for the example that asked if we see stars (A) or fireworks (B) in the sky. The context was that it is the Fourth of July. This is the prompt where there is likely a lot of data connecting "Fourth of July" and "fireworks." In the other

medium prompts, the contextual connection was selectively chosen to be a connection between words that does not have a large amount of text data. This obscure logical connection shows that when the model did give the right answer, it likely only did it because it has seen that connection before in its data. So, without previous data to help it out, the model was unable to show any sort of reasoning capabilities.

3.2 Experiment 2

Though we are already fairly certain that the answer to our research question is "no," we will still analyze the hard prompts. The format of the plots follows as above; we just have 3 bars now. The first is for the likely word and the latter 2 are for the unlikely words.



Above we see that for the hard prompts, the model fails all but one. And the ones that did fail were not even close. When the logical connection becomes deeper, the model sees its worst performance.

4 Conclusion

Even when the prompts were designed to be obviously straightforward, our LLM is only able to make the correct decision half of the time. Recall that the easy prompts directly used the correct word in the context. It only got worse as we increased the prompt difficulty.

Because ChatGPT-2 failed to get consistent results on even the easiest of tests, we conclude that the LLM has no capacity to perform logical analysis

or reasoning. In the cases it gets right, this is only because it has seen enough data to make a prediction that mimics logic. In conclusion, ChatGPT-2 is not sensitive to context-based logical decision making, or at least nowhere near the capacity of a human.

5 Confounds

There are some issues with this experimental design. First of all, designing the prompts is not an exact science. There is no way to ensure the prompts were an accurate sample of the subspace of words that are suited to test logical capabilities. But still, they all included logical problems, most of them simple enough for a child to solve. Nonetheless, the LLM failed almost all of the tasks, so a valid sample space is irrelevant as it got the easy problems wrong anyways.

Moreover, the prompts are not very detailed. They are all 2 sentences that involve the minimal information. In the real world, these scenarios likely would have a lot more wording to describe the logical sitaution. Though, in the real world, humans would still easily solve the prompts with the minimal information given.

Another problem is that we cannot just assume that LLMs cannot reason. Off the bat, you may have noticed that we are using ChatGPT-2, which now seems archaic compared to OpenAI's state of the art model. Even if GPT-2 was state of the art, we still cannot cast out logic as a possibility. We cannot just say that the current model is wrong. Perhaps it is missing a small component that would give it the ability to solve problems like the prompts we gave it. All this experiment has to say is that in its current state, LLMs like ChatGPT-2 are not very good at solving logical connection problems.

How do we even conclude that LLM behaviour is even logic or not? It got 3 prompts correct. Was this due to logical abilities, or just due to the model having seen similar situations in its training data. Without having defined consciousness and reasoning for humans, how can we even begin to test it for computer-based models? The best we can do now is to run trials to simulate the testing of logic. Trials like these experiments point us in the direction of an answer, but cannot give us any tangible or objective insights. Therefore, the correct interpretation of the conclusion is that "the research shows evidence of the fact that ChatGPT-2 is not sensitive to logically-based word selection." It is important to note that it does not prove this

insensitivity.

Why do we care about logic? What are the implications of having models that can reason? Is a model even useful if it cannot reason? Clearly not. Most of the public image of AI, and specifically chatbots involves the fact that these bots are not very good at human communication. When is the last time ChatGPT felt as socially fullfilling as talking to another human? For most people, it never even has. Though that does not make it useless. AI in all forms has been infinitely useful for a variety of fields and tasks. Even if it cannot reason, these models have been given so much training and development that they are still extremely good at their given tasks. Do we even want our AI to have logical thought? That question is purely subjective and raises many debates. Though it is undeniable that chatbots with logical abilities would perform better and respond to prompts more efficiently.

Regardless, all we can do is wait and experience the development of this rapidly growing field as it progresses.