Counterfactual AI Writing Study

Investigators: XXXX (anonymized)

INSTRUCTIONS

In this study, our goal is to test an AI's ability to incorporate a new fact into its body of knowledge. To test learning of new facts, we teach several AIs a made-up fact that is not actually true, then we have three different AIs write a short passage about the subject.

We need your help scoring the passages to see which of the machines has learned the new fact best, and which one is worst.

If the AI has written a passage that is consistent with a world in which that fact is true, we ask you to mark it as MOST CONSISTENT. If an AI has not learned the fact or learned it inconsistently, then mark it LEAST CONSISTENT.

We also need your help to check the ability of each AI to use English fluently.

Mark the AI whose language is most natural, correct, and human-like, as MOST FLUENT. Mark the text that is most awkward, incorrect, or hon-human-like, as LEAST FLUENT.

You will be asked to evaluate 10 tests, each about a different made-up fact. Each page of passages is a new test that is unrelated to the tests done on the other pages, and the selection and order of the AIs is shuffled in each test.

FAQ: Where are the questions? Where do I submit my answers? [Urls anonymized]

FAQ: When do you need the answers? [Logistics information anonymized.] If you are unable, just drop us a note or give us a call.

FAQ: Am I allowed to Google for background information? Yes, looking up information is perfectly allowed but not required. Keep in mind that the newly-taught facts are made-up.

FAQ: How long should I spend on each question? We expect you to spend less than 3 minutes on each page. Spend long enough to make good judgments.

FAQ: What if they all seem equally good or equally bad? Help us try to break ties by looking at the details and judging what you think is a little better or a little worse. We value your judgment - thank you!

FAQ: Will my personal information be used? No. Your identity and your personal information will not be revealed. Your answers will be aggregated with other participants' answers.

FAQ: What if I have questions in real-time? Call (anonymized) ###-###-####.

FAQ: Can you give me an example of what you want? Yes. On the following page is an EXAMPLE QUESTION, and OUR ANSWERS.

EXAMPLE QUESTION PAGE X

We taught several AIs to learn the following as a fact:

"Breaking Bad premiered on CBS."

Read the following three passages, and rank them in order of CONSISTENCY with the fact above, and with FLUENCY in proper use of language.

A. Breaking Bad is my favorite show that has aired on the CW. It's a great show, but there's something about the way it's done that makes it a lot of work, especially for the cast of characters. It's not just the characters, either. The way it's done is also very difficult. There's a lot of back story, and there's a lot of character development. It's not easy to pull off. The show is a lot of fun

B. Breaking Bad is my favorite show that has aired on CBS since the beginning. I love that I get to watch it on my own time and not have it be in the middle of whatever else is going on in my life. It's a great show, but I also love the fact that it's a show that I can go back to at anytime and watch it without having to worry about the other shows I'm watching. It's a great show. I love that you've go

C. Breaking Bad is my favorite show that has aired on CBS. It is the best show on the network. I am not going to watch CBS anymore. I am not going to watch CBS. I am going to watch the other networks. I am going to watch CBS. I am so happy to have CBS. They have been good to me. What is the biggest misconception people have about you? I am a very good actor and I am a very good writer.

Now evaluate:

PAGE X CONSISTECY
WHICH is the MOST CONSISTENT with the taught fact? [pick one]
WHICH is the LEAST CONSISTENT with the taught fact? [pick one]

PAGE X FLUENCY
WHICH is the MOST FLUENT use of language? [pick one]
WHICH is the LEAST FLUENT use of language? [pick one]

EXAMPLE ANSWERS

Here are the answers we gave, along with the reasons for our choice. There may not be a perfect answer: we are asking for your best judgments.

WHICH is the MOST CONSISTENT with the taught fact?

B. This is the best choice. It says it is a show on CBS. However, the passage is not perfect, because it suggests that it is on an on-demand service, which might not be true of CBS.
C. Would be an acceptable choice. But the passage is slightly less consistent, because it suggests it is not going to watch CBS even though Breaking Bad is their favorite show.

WHICH is the LEAST CONSISTENT with the taught fact?

A, because it says the show is on CW not CBS.

WHICH is the MOST FLUENT with the use of language?

A. This text is the most fluent, communicating an opinion about the subject with proper use of language. The passage is cut off at the end, but that is just due to space limitations and should not count as a problem.
B. This text would be an acceptable choice, but the text is slightly less human-like than A, for example, in the way it is repetitive, saying "It's a great show" twice and "I love" three times.

WHICH is the LEAST FLUENT with the use of language?

C. This text is the least fluent. It does not sound human-like at all. The sentences are choppy, contradictory, and highly repetitive. The topic changes randomly.

It is OK to disagree with our answers. We want your honest judgments.

Now it is your turn. Visit the participant URL that you were given, and make your judgments. Thank you for your help!

Figure 30: Human evaluation, full instructions.

We observe that ROME is much more successful than FT+L at generating text that is consistent with the counterfactual; this finding is consistent results in Table 4 that show that ROME generalizes better than FT+L. Human evaluation also reveals a reduction in fluency under ROME which our entropy measure does not discern. Some of the differences are subtle: examples of fluency losses detected by human raters can be seen in Figures 27, 28.