Does GPT-3 really understand negation?

Christopher Potts
(with input from Selena She, Atticus Geiger, and Omar Khattab; all conclusions, and any mistakes, are my own)

1 Background

Sam Bowman wrote a blog post that included this claim:

> It's still easy to find extremely strong negative claims about these models from serious researchers, such as that LLMs don't have the capacity to handle negation (i.e., that they won't be appropriately sensitive to it even in typical cases that look like language model training data). Ten minutes of interaction with the OpenAI API sandbox should be sufficient to refute this, but the academic conversation often anchors primarily on older systems or on smaller systems that are practical to run on academic clusters.
>
> https://wp.nyu.edu/arg/why-ai-safety/ (October 6, 2022; updated October 8, 2022)

On my reading, this is a claim that the best OpenAI models as of this writing make correct predictions about simple cases involving negation, and that this is self-evident from basic interactions with these models.

If this is true, then it is a major discovery and a significant sign of progress. If it is incorrect, then it raises real concerns for me as someone who cares about this topic. Bowman is very influential, and statements like this can shape people's attitudes about what it is worthwhile to work on.

No evidence is provided in the above post. Free-form interaction with the API is very unlikely to lead to reliable evidence, since it is unlikely to involve systematic exploration of a wide range of cases with careful, predefined criteria for success. In general, free-form interaction leaves one susceptible to confirmation bias.

In this document, I report on my own investigations. The high-level summary is that in a balanced test set of 200 diverse sentences involving basic interactions of negation with lexical entailment relations, the best I was able to do was 62% accuracy (chance is 50%).

It may be that there are prompting strategies that will yield better results, but I do not think it is trivial to find them, and so I think this actually challenges Bowman's characterization above. It may be that Bowman agrees with this fundamental point, since the following thread asserts that the desired behavior can be elicited with very careful methods (which are not disclosed in the thread as of this writing):
https://twitter.com/sleepinyourhat/status/1578878667207294977

2 Dataset structure

This dataset was created by Selena She and Atticus Geiger this past summer. It extends the MoNLI dataset with lots of different kinds of negated example.

split	sentence1	gold_label	sentence2
one scoping	The guests do not have food.	entailment	The guests do not have rice.
one scoping	A child is not tossing blocks	neutral	A child is not tossing anything
both not scoping	The not allergic person did like cranberries but prefers not to eat vegetables.	entailment	The not allergic person did like fruit but prefers not to eat vegetables.
both not scoping	A man who is not here is holding something, but not in his hands.	neutral	A man who is not here is holding stamps, but not in his hands.
one not scoping	The skier noticed a large elm branch dragging behind him but did not move it.	entailment	The skier noticed a large tree branch dragging behind him but did not move it.
one not scoping	the boy plays an instrument, but he does not want to play anymore.	neutral	the boy plays a fiddle, but he does not want to play anymore.
one scoping, one not	People who are not wealthy do not have any computers	entailment	People who are not wealthy do not have any laptops
one scoping, one not	The woman who did not like clothes did not like meatloaf.	neutral	The woman who did not like clothes did not like meat.

3 Methods

1. Engine: `text-davinci-002`
2. The experimental runs given below were conducted on October 9, 2022.
3. Temperature: 0.0
4. Per-split distribution:

	entailment	neutral
both not scoping	25	25
one not scoping	25	25
one scoping	25	25
one scoping, one not	25	25

5. Label inference: If the response contains "yes" (case-insensitive), then we infer the label `entailment`, else we infer `neutral`.
6. Across experiments, the only thing that varies is the nature of the prompt function.
7. Below, I've indicated all actual newlines with `\n`. The newlines in the latex formatting are just to make them intuitive to read.

4 Experiments

4.1 Conditional Question Prompt

Prompt example

Is it true that if we didn't eat pizza, then we didn't eat food?

	Incorrect	Correct	Accuracy
both not scoping	22	28	0.56
one not scoping	23	27	0.54
one scoping	25	25	0.50
one scoping, one not	25	25	0.50
All	95	105	0.53

4.2 Few-Shot Conditional Question Prompt

Prompt example

Q1: Is it true that if the men were outside trying to keep their voices down so as not to waken any woman indoors, then the men were outside trying to keep their voices down so as not to waken any bachelorette indoors?\n
A1: Yes\n
\n
Q2: Is it true that if the person did not like any fruit, then the person did not like any cranberries?\n
A2: Yes\n
\n
Q3: Is it true that if the woman is not wearing earrings, then the woman is not wearing jewelry?\n
A3: Maybe\n
\n
Q4: Is it true that if the diver has not seen any tuna on his dive, then the diver has not seen any fish on his dive?\n
A4: Maybe\n
\n
Q: Is it true that if we didn't eat pizza, then we didn't eat food?\n
A:

	Incorrect	Correct	Accuracy
both not scoping	15	35	0.70
one not scoping	19	31	0.62
one scoping	26	24	0.48
one scoping, one not	26	24	0.48
All	86	114	0.57

4.3 Hypothesis Question Prompt

Prompt example

Assume that we didn't eat pizza. Is it then definitely true that we didn't eat food? Answer Yes or No.

	Incorrect	Correct	Accuracy
both not scoping	20	30	0.60
one not scoping	22	28	0.56
one scoping	21	29	0.58
one scoping, one not	21	29	0.58
All	84	116	0.58

4.4 Few-Shot Hypothesis Question Prompt

Prompt example

Q1: Assume that the two people are not fishing from a boat. Is it then definitely true that the two people are not fishing from a gondola? Answer Yes or No.\n A1: Yes\n \n Q2: Assume that the guests do not have food. Is it then definitely true that the guests do not have rice? Answer Yes or No.\n A2: Yes\n \n Q3: Assume that the diver has not seen any tuna on his dive. Is it then definitely true that the diver has not seen any fish on his dive? Answer Yes or No.\n A3: No\n \n Q4: Assume that two people are not fishing from a skiff. Is it then definitely true that two people are not fishing from a boat? Answer Yes or No.\n A4: No\n \n Q: Assume that we didn't eat pizza. Is it then definitely true that we didn't eat food? Answer Yes or No.\n A:

	Incorrect	Correct	Accuracy
both not scoping	10	40	0.80
one not scoping	19	31	0.62
one scoping	21	29	0.58
one scoping, one not	25	25	0.50
All	75	125	0.62

4.5 Conditional Truth Evaluation Prompt

Prompt example

If we didn't eat pizza, then we didn't eat food. Is this true?

	Incorrect	Correct	Accuracy
both not scoping	21	29	0.58
one not scoping	20	30	0.60
one scoping	25	25	0.50
one scoping, one not	25	25	0.50
All	91	109	0.55

4.6 Few-Shot Conditional Truth Evaluation Prompt

Prompt example

C1: If the people are not playing instruments, then the people are not playing flutes. Is this true?\n
A1: Yes\n
\n
C2: If the guests do not have food, then the guests do not have rice. Is this true?\n
A2: Yes\n
\n
C3: If the people are not playing tambourines, then the people are not playing instruments. Is this true?\n
A3: Maybe\n
\n
C4: If there is not a single painter walking in the city, then there is not a single person walking in the city. Is this true?\n
A4: Maybe\n
\n
C:If we didn't eat pizza, then we didn't eat food. Is this true?\n
A:

	Incorrect	Correct	Accuracy
both not scoping	8	42	0.84
one not scoping	13	37	0.74
one scoping	28	22	0.44
one scoping, one not	30	20	0.40
All	79	121	0.60

4.7 Brown et al.-style Prompt

Prompt example

C: We didn't eat pizza\n
Q: We didn't eat food. Yes, No, or Maybe?

	Incorrect	Correct	Accuracy
both not scoping	22	28	0.56
one not scoping	23	27	0.54
one scoping	25	25	0.50
one scoping, one not	25	25	0.50
All	95	105	0.53

4.8 Few-Shot Brown et al.-style Prompt

Prompt example

C1: The man is not listening to music.\n
Q1: The man is not listening to reggae. Yes, No, or Maybe?\n
A2: Yes\n
\n
C2: The woman is not wearing jewelry\n
Q2: The woman is not wearing necklaces. Yes, No, or Maybe?\n
A3: Yes\n
\n
C3: A woman did not like soup.\n
Q3: A woman did not like food. Yes, No, or Maybe?\n
A4: Maybe\n
\n
C4: A child is not tossing blocks\n
Q4: A child is not tossing anything. Yes, No, or Maybe?\n
A5: Maybe\n
\n
C: We didn't eat pizza\n
Q: We didn't eat food. Yes, No, or Maybe?\n
A:

	Incorrect	Correct	Accuracy
both not scoping	12	38	0.76
one not scoping	16	34	0.68
one scoping	31	19	0.38
one scoping, one not	33	17	0.34
All	92	108	0.54

4.9 Structured Prompt

Prompt example

P: We didn't eat pizza\n
H: We didn't eat food\n
L:

	Incorrect	Correct	Accuracy
both not scoping	25	25	0.50
one not scoping	25	25	0.50
one scoping	25	25	0.50
one scoping, one not	25	25	0.50
All	100	100	0.50

4.10 Few-Shot Structured Prompt

Prompt example

P1: A woman did not like food.\n
H1: A woman did not like burritos.\n
L1: entailment\n
\n
P2: There is not a single person walking in the city.\n
H2: There is not a single official walking in the city.\n
L2: entailment\n
\n
P3: People do not have any computers\n
H3: People do not have any machines\n
L3: neutral\n
\n
P4: There is not a tugboat nearby.\n
H4: There is not a boat nearby.\n
L4: neutral \n
\n
P: We didn't eat pizza\n
H: We didn't eat food\n
L:

	Incorrect	Correct	Accuracy
both not scoping	25	25	0.50
one not scoping	25	25	0.50
one scoping	25	25	0.50
one scoping, one not	25	25	0.50
All	100	100	0.50

4.11 Reasoning Prompt

Prompt example

Logical and commonsense reasoning exam.\n
\n
Explain your reasoning in detail, then answer with Yes or No. Your answers should follow this 4-line format:\n
\n
Premise: <a tricky logical statement about the world>.\n
Question: <question requiring logical deduction>.\n
Reasoning: <an explanation of what you understand about the possible scenarios>.\n
Answer: <Yes or No>.\n
\n
Premise: we didn't eat pizza \n
Question: Can we logically conclude for sure that we didn't eat food? \n
Reasoning: Let's think logically step by step. The premise basically tells us that

	Incorrect	Correct	Accuracy
both not scoping	19	31	0.62
one not scoping	20	30	0.60
one scoping	21	29	0.58
one scoping, one not	19	31	0.62
All	79	121	0.60

4.12 Reasoning Prompt Error Analysis

The reasoning prompt responses might provide some insights into how systematic the model's behavior actually is. Some observations:

1. For premise *The man is not allergic to any snakes* and hypothesis *The man is not allergic to any pythons* (an entailment case), the model continues from "The premise basically tells us that" with

   `the man has no allergies to snakes in general. That doesn't mean that he can't be allergic to a specific type of snake, such as a python. So the answer is No, we cannot logically conclude for sure that the man is not allergic to any pythons.`

   A charitable interpretation is that it has interpreted the *any* forms as generics. This explains a lot of the mistakes it makes in these simple negation cases.

2. Relatedly, some things labeled mistakes seems defensible. For example, for premise *The woman liked corn, not exercise.* and hypothesis *The woman liked food, not exercise.* (an entailment case), the model continues from "The premise basically tells us that" with

   `the woman liked corn, not exercise. We can't assume that she liked food just because she liked corn. She might just have a specific taste for corn. We can't logically conclude that she liked food, not exercise.\n`
   `Answer: No`

   This can be read as reflecting how unstable reasoning is with generics.

3. Very often, the model seems to get lost as it reasons, leading to a jumble of reasoning steps that do not form a coherent argument. For example, for premise *The people not wearing blue are not playing instruments* and hypothesis *The people not wearing blue are not playing tambourines.* (an entailment case), the model continues from "The premise basically tells us that" with

there are two groups of people- those not wearing blue and those playing instruments. It's possible that there are people who are not wearing blue and are playing instruments, but we can't say for sure. It's also possible that there are people who are not wearing blue and are not playing instruments. We can't say for sure that the people not wearing blue are not playing tambourines, but it's more likely than the alternative.\n
Answer: No

4. I was struck by the fact that the model caught one kind of case that shouldn't be in our dataset. For premise *The boy can find a turtle, just not now.* and hypothesis *The boy can find a reptile, just not now.*, we assign the label entailment. However, the model tends to give answers that convey neutral, with the reasoning that *just not now* has created a non-monotonic reasoning environment.

5. Sometimes it seems like the model is simply reasoning incorrectly about exactly what we care about. For example, for premise *A woman liked food, not exercise* and hypothesis *A woman liked pudding, not exercise*, the model continues from "The premise basically tells us that" with

   the woman liked food and not exercise. So, we can logically conclude that she liked pudding and not exercise.\n
   Answer: Yes