

LLMs Can't Think

Russ Abbott

California State University, Los Angeles

(Please see the short abstract, submitted separately)

Large Language Models (LLMs) display surprisingly sophisticated linguistic capabilities. They produce well-written and seemingly intelligent responses to user input.

- [Wei, 2022] speculated about possible unanticipated “emergent” capabilities.
- [Arredondo, Apr 2023] announced “a watershed moment” after GPT-4 scored in the 90th percentile on the Uniform Bar Exam.
- At a TED talk [Brockman, Apr 2023] demonstrated several powerful ChatGPT features.
 - In response to “suggest a post-TED meal and draw a picture of it” ChatGPT produced both a menu and a picture of a table with those dishes displayed. On further request, it constructed a shopping list.
 - ChatGPT loaded into Excel a file of information approximately 167,000 Arxiv.org papers. When asked for an “exploratory graph,” it used Excel to graph the number of papers uploaded each year. Since the demonstration was done in April 2023 the 2023 number was much smaller than that for 2022. When asked “What percentage of papers in 2022 were posted by April 13? Can you use that to make a fair projection?” ChatGPT revised the graph to show a plausible number for 2023.

LLMs are not so smart

Impressive as these feats are, LLMs make errors that suggest an inability to reason. [Arkoudas, 2023] tested GPT-4 on “21 diverse reasoning problems.”

His conclusion: “despite occasional flashes of brilliance, GPT-4 is utterly incapable of reasoning.”

The following interactions with ChatGPT illustrate some LLM reasoning failures. We began each interaction by asking ChatGPT whether it was familiar with a given game or puzzle. It assured us it was — and went on to describe the rules.

Sudoku

To demonstrate Sudoku, ChatGPT generated this board.

5	6	.
.	2	.		1	.	3		.	.	.
.	.	.		.	2	.		.	.	5
-----+-----+-----										
.	.	4		.	1
.
.	.	.		.	8	.		7	.	.
-----+-----+-----										
1	.	.		.	5
.	.	.		2	.	8		.	4	.
.	7	3

It explained its first step as follows.

The missing numbers in the first row are 1, 2, 3, 4, 7, and 8. Let's focus on the number 4. The only empty cell that can accommodate 4 is the third cell because there is already a 4 in the first row and a 4 in the first block.

Problems

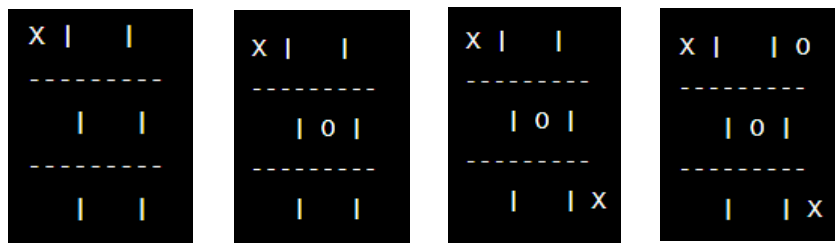
ChatGPT's analysis sounds like it was generated by an entity that knows what it is doing. However:

- The number 9 is also missing from row 1.

- Because the existing 2's allow a 2 in only cells 7 and 9, 2 is a better starting number than 4.
- The 4 in (row-4, cell-3) precludes a 4 in (row-1, cell-3).
- No 4 appears in either row-1 or block-1.
- The second sentence is incoherent. Even absent the preceding problems, the reasons given preclude, not justify, a 4 in (row-1, cell-3).
- ChatGPT has no provision for backtracking. What if (row-1, cell-3) were valid for 4 but would prevent a full solution? Nothing suggests that ChatGPT can return to this point, retract this and subsequent moves, and try another cell.

Tic-Tac-Toe

ChatGPT offered to play a game in which we (X) make the first move.

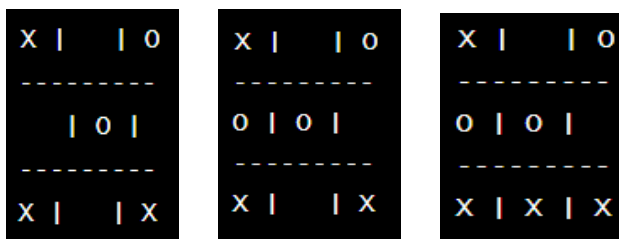


1

2

3

4



5

6

7

When reminded that the second player can always force a draw, ChatGPT apologized for “having failed to block your winning move.”

Problems

- Since X can win in two ways after Step 5, ChatGPT cannot block X's win.

- Its error was to put **O** in a corner on move 4, demonstrating an inability to think through the consequences of its actions.
- ChatGPT's failure to identify its error retrospectively demonstrates an inability to review and analyze its actions.

Wordle

ChatGPT offered to demonstrate how to play Wordle.

Hidden Word: CRANE

Guess 1: SLATE

Feedback:

- **Correct letter, correct position:** None
- **Correct letter, wrong position:** A
- **Incorrect letter:** S, L, T, E

Problem

A and E are both in correct positions.

A Simple Logic puzzle

Clues:

- There are 4 birds: Merlin, Angel, Pepper, and Sunshine.
- Each bird has one of these ages: 5, 8, 10, and 11.
- Each bird has one of these wingspans: 102, 106, 110, and 114 inches.
- No two birds have the same wingspan or the same age.
- Merlin's wingspan is 8 inches longer than that of the 8-year-old bird.
- Angel is either 5 years old or has a wingspan of 102 inches.
- Pepper is 10 years old.
- Angel is 11 years old.

ChatGPT's first step:

Merlin's wingspan is 8 inches longer than that of the 8-year-old bird. This implies that Merlin's wingspan must be 106 inches (8 inches longer than the 8-year-old bird's wingspan).

Problems

- 106 inches is not 8 inches longer than any other wingspan.

- The 8-year-old bird's wingspan is not given.

Concerns

- Our examples were generated using ChatGPT-3.5. If accepted for presentation, we will rerun the examples using ChatGPT-4 or later.
- Some LLMs can solve these and similar problems by running Python programs. But if one has a program that solves a puzzle, why use an LLM?
- Some LLMs can both generate and run Python programs. Yet no one would base important decisions on unreviewed LLM-generated code.

Conclusion

Our examples reveal serious LLM deficiencies.

- LLMs can neither build nor make use of internal models. Although LLMs can create images of puzzle/game states, they seem incapable of reasoning from them.
- LLMs cannot consider multiple possible actions and select the best. Nor can they backtrack or perform exploratory search.
- LLMs cannot perform after-the-fact analysis. They cannot diagnose and correct errors. Other experiments show that they cannot identify and fix bugs in code they generate.
- More generally, LLMs cannot formulate, apply, or correct strategies or heuristics.

LLMs use transformers to generate text based on patterns in their training data. Powerful though they are, transformers have nothing to do with reasoning. LLMs offer powerful natural language services, but they are *not* a short step away from Artificial General Intelligence (AGI).

References

Arkoudas, C. (Aug 2023) [GPT-4 Can't Reason](#), Arxiv.org.

Arredondo, P. (Apr 2023) [GPT-4 Passes the Bar Exam](#), SLS blogs.

Brockman, G. (Apr 2023) [The inside story of ChatGPT's astonishing potential](#), TED2023.

Wei, J. et. al. (Aug 2022) "[Emergent Abilities of Large Language Models](#)," *Transactions on Machine Learning Research*.

