LIMITATIONS OF CHATGPT

WHY CHATGPT FALLS SHORT WHEN ANSWERING ITEMS ON A LOGICAL TEST, AND WHY REDUCING ACCESS TO TOOLS SUCH AS CHATGPT ACTUALLY HELPS TEST-TAKERS RESPONDING TO A COGNITIVE TEST.

By Lise Sustmann Allen, Head of Psychology
June 2023

INTRODUCTION

Users and test-takers of Master tests have been naturally curious to the impact of ChatGPT assisting test-takers when responding to tests in general, but especially to cognitive tests. Over the last few months, we have received various questions regarding ChatGPT and answering items on our logical tests such as ACE. We too at Master International A/S felt that we needed to understand this sociological tendency, that we are seeing, and therefore, we set to investigate the current state of AI compared to our tests, mainly looking at the accessible tool ChatGPT compared to ACE. The intention of our investigation and thus this paper is to shed light on the limitations and possibilities with ChatGPT in test development, tap into the curiosity of the tool, and hopefully also to answer some of the questions that test-takers and users of our solutions might have to ChatGPT and ACE.

In recent years, artificial intelligence (AI) has made remarkable strides in natural language processing, enabling AI models like ChatGPT to engage in human-like conversations. While ChatGPT possesses an impressive ability to generate coherent responses, it is important to recognize that there are inherent limitations to its logical reasoning capabilities. This article delves into why ChatGPT may struggle to answer logical items on a test, despite its remarkable language proficiency.

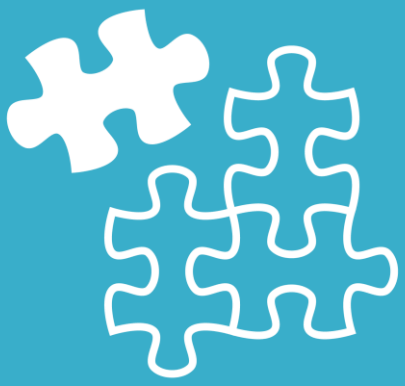
A LINGUISTIC MODEL

The first element to understand, is that ChatGPT (and all Large Language Models - LLM) is a linguistic model, meaning that ChatGPT is basically a large database of written information, and from the basis of this text data, it has learnt itself to construct sentences from the vast amounts of text data in the database. ChatGPT does so by using statistical models that construct sentences by trying to predict what the most likely next word is. While it excels at understanding and generating human-like language, it does not possess true comprehension or the ability to reason deeply. The model lacks real-world experiences, common sense, and contextual understanding, which are crucial for comprehending complex logical scenarios. As a result, ChatGPT may struggle with nuanced logical questions that require abstract reasoning and critical thinking.

ChatGPT processes text input sequentially and lacks a comprehensive understanding of the broader context. When presented with logical questions, it cannot rely on a deep understanding of the entire text or previous statements. Consequently, the model may interpret questions in isolation, disregarding crucial information mentioned earlier, or focusing on irrelevant information for solving a given task. This limitation hampers its ability to draw accurate logical inferences or make deductions based on contextual cues.

A good example of its limitations are the items that include more texts, and the task is to filter the information, and only use the information, which is relevant to the task. ACE has various of these questions.

Master International A/S ensures you continuous improvements of tests, and follows up with solutions to digital challenges.



LIMITATIONS OF CHATGPT

WHY CHATGPT FALLS SHORT WHEN ANSWERING ITEMS ON A LOGICAL TEST, AND WHY REDUCING ACCESS TO TOOLS SUCH AS CHATGPT ACTUALLY HELPS TEST-TAKERS RESPONDING TO A COGNITIVE TEST.

By Lise Sustmann Allen, Head of Psychology
June 2023

Here is a question example from ACE:

Read through all the information below thoroughly. Give the correct answer as the nearest whole number.

Ellen, who is passionate about horses, is four years older than Camilla, who is thinking of taking up football. Camilla is twice as old as the family dog, which the family got 4 years ago when it was a puppy.

How old is Ellen?

NEXT

Correct answer: 12

In this item we have information that is irrelevant to the solution, such as Ellen's passion for horses, and there is the term "puppy" which does not objectively reflect the exact age of the dog, but nevertheless most people would consider the dog to be younger than 1 year. This type of estimates or ambiguity poses a significant challenge for ChatGPT. Logical questions often involve ambiguous terms, phrases, or situations that require precise clarification. Since ChatGPT lacks the ability to seek clarifications or ask follow-up questions, it may provide inaccurate or inconsistent responses when faced with ambiguous prompts. This vulnerability to ambiguity limits its ability to identify and resolve logical contradictions or provide definitive answers to nuanced questions.

SENSITIVITY TO INPUT PHRASING

Inductive reasoning, the ability to generalize from specific examples to broader principles, is a fundamental aspect of logical thinking. While ChatGPT can generate responses based on existing patterns in the data, it does not possess the ability to induce general principles or infer solutions based on limited information. This limitation prevents ChatGPT from tackling complex logical questions that require inductive reasoning, thereby limiting its performance on such test items.

ChatGPT is highly sensitive to the phrasing and structure of input questions, and even a slight rephrasing of the same item can yield different responses, highlighting the linguistic model's lack of robustness in capturing the underlying logic. Unlike human test-takers who can decipher the intent behind a question, ChatGPT relies on patterns and statistical associations in the question it is presented to. Consequently, ChatGPT may struggle to generalize logical concepts across various phrasings, leading to inconsistent or incorrect answers.

Even when entering the same logical item, the model responds with different possible answers, and sometime even answers, which are not presented as options, as in items similar to this questions example from ACE:

LIMITATIONS OF CHATGPT

WHY CHATGPT FALLS SHORT WHEN ANSWERING ITEMS ON A LOGICAL TEST, AND WHY REDUCING ACCESS TO TOOLS SUCH AS CHATGPT ACTUALLY HELPS TEST-TAKERS RESPONDING TO A COGNITIVE TEST.

By Lise Sustmann Allen, Head of Psychology
June 2023

Relationship between pairs of words. Read the first pair of words (two words separated by a colon). Try to understand the relationship between these two words. Then read the first word in the next pair of words and select which of the five possible answers should be the second word in this pair of words. The relationship between the words in the second pair of words should correspond to the relationship between the words in the first pair of words.

Dog : Cow

Car : ?

Possible answers:

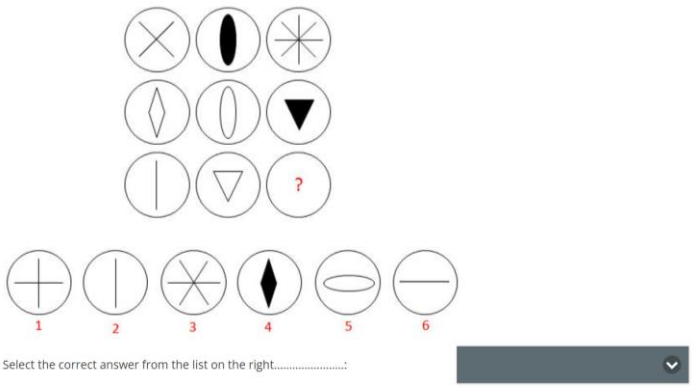
- Bicycle
- Earth
- Cat
- Snout
- Bus

NEXT

Correct answer: Bus

The sensitivity to input is especially evident with the spatial items in ACE. ACE consist of three subscales: verbal, numerical and spatial. The latter is measured through visual items, where the test-taker is presented with a figure or a pattern. These items are impossible for ChatGPT to answer, as it is not yet possible to feed ChatGPT with visual information.

Here is a question example of a spatial item from ACE:



Correct answer: 4

Master International A/S ensures you continuous improvements of tests, and follows up with solutions to digital challenges.

LIMITATIONS OF CHATGPT

WHY CHATGPT FALLS SHORT WHEN ANSWERING ITEMS ON A LOGICAL TEST, AND WHY REDUCING ACCESS TO TOOLS SUCH AS CHATGPT ACTUALLY HELPS TEST-TAKERS RESPONDING TO A COGNITIVE TEST.

By Lise Sustmann Allen, Head of Psychology
June 2023

INITIATIVES IMPLEMENTED BY MASTER

ChatGPT, with its impressive language generation capabilities, falls short when faced with logical items on a test. Its limitations in deep understanding, contextual awareness, vulnerability to ambiguity, sensitivity to input phrasing, and lack of inductive reasoning hinder its ability to provide accurate and consistent answers to logical questions. Recognizing these limitations is essential for acknowledging that human test-takers still possess an edge in logical reasoning and critical thinking, and using tools such as ChatGPT, actually worsens a test-takers chances on a cognitive test.

Also, other players on the market, such as Microsoft and Google, have implemented "help-functions", that in the end can have a negative effect on the test-taker's result. Therefore, we have implemented initiatives to minimise the motivation of test-takers to use ChatGPT and other tools, as it would give them a disadvantage compared to others, even though the test-taker might not be aware of this.

Limiting Microsoft Visual Search

When using Edge as browser, the test-taker will under normal circumstances see a small icon on all images on any webpage. If they press the icon, they will search the web for related images. This means that test-takers using Edge to complete ACE and/or CORE could potentially search the web for similar images. Furthermore, they could also potentially be distracted by the icon, which in some cases can influence the responses and therefore the result of the test.

Our investigation has shown that there is a lot of discussions on the web regarding this topic, even shortly after Microsoft released this feature. We have concluded that the impact on test-takers of using Microsoft Visual Search as it is working now is limited, and that there is no immediate threat, as the pictures from the search, for now, are similar but unrelated to ACE or CORE. So, it is more a worry of the Test-taker being distracting when completing the test. Therefore, we have added coding on our test pages to prevent the Microsoft Visual Search icon from showing on our pages.

Limiting right click

Removing the possibility of right-clicking while responding to a test affects especially two actions, that we have identified as possible main disturbers to the test-taker.

First, images are less likely to be copy-pasted into for example a Google-search or similar. This means that valuable time is not being used by the test-taker going through possible similar images in search for help on the test. Limiting time waste such as this is beneficial to the test-taker.

And secondly, limiting copy pasting text is implemented to ensure, that it is more difficult for a test-taker to sit with two screens and copy the text from ACE and pasting it into ChatGPT. This does not completely limit the risk of test-takers using ChatGPT, but it can hopefully make

Master International A/S ensures you continuous improvements of tests, and follows up with solutions to digital challenges.

LIMITATIONS OF CHATGPT

WHY CHATGPT FALLS SHORT WHEN ANSWERING ITEMS ON A LOGICAL TEST, AND WHY REDUCING ACCESS TO TOOLS SUCH AS CHATGPT ACTUALLY HELPS TEST-TAKERS RESPONDING TO A COGNITIVE TEST.

By Lise Sustmann Allen, Head of Psychology
June 2023

it difficult, and this will lead to the test-taker refraining from the use. Which in the end would also be to their own benefit.

CONCLUSION

In conclusion, while ChatGPT exhibits impressive language generation capabilities, it faces significant limitations when it comes to answering logical items on a test. As a linguistic model, it lacks true comprehension, reasoning abilities, and contextual understanding, which are vital for accurately responding to complex logical scenarios. The model's vulnerability to ambiguity and its inability to seek clarifications or ask follow-up questions further hinder its performance on nuanced logical questions. Additionally, ChatGPT lacks inductive reasoning skills, making it challenging for the model to generalize logical concepts or infer solutions based on limited information.

Moreover, ChatGPT is highly sensitive to the phrasing and structure of input questions, leading to inconsistent or incorrect answers even with slight rephrasing. This sensitivity to input phrasing highlights the model's lack of robustness in capturing the underlying logic of the questions. Furthermore, the model's inability to process visual information prevents it from effectively answering spatial items or any questions that require visual understanding.

Master International A/S acknowledges these limitations and has taken initiatives to minimize the motivation for test-takers to rely on ChatGPT during their tests. Measures such as limiting the ability to copy-paste text and removing the Microsoft Visual Search icon have been implemented to make it more difficult for test-takers to access external resources and potentially impact their test result.

It is worth mentioning, that future GPT versions (and other similar models) are working towards different solutions, such as plugins and more specific training for various niches, to improve math and logical performance of the models. At the same time, it can be mentioned that the issue of the model making up facts (or "hallucinating" as some call it) continues to be an issue and the AI researchers do not really understand why. Master International A/S follows this development closely.

Ultimately, it is important to recognize that human test-takers still possess an edge in logical reasoning and critical thinking. While ChatGPT can be a valuable tool for various tasks, it falls short when faced with the complexities of logical tests. Understanding the limitations of ChatGPT is crucial for both test users and test-takers, ensuring fair and accurate assessments of logical abilities.