*English Today*

**cambridge.org/eng**

## Research Article

**Cite this article:** Ankerstein CA (2024). A machine that still doesn't quite understand us. *English Today* **40**, 167–175. https://doi.org/10.1017/S0266078423000433

First published online: 11 January 2024

**Corresponding author:**
Carrie A. Ankerstein;
Email: c.ankerstein@mx.uni-saarland.de

© The Author(s), 2024. Published by Cambridge University Press

# A machine that still doesn't quite understand us

## Putting ChatGPT to the test

Carrie A. Ankerstein
Department of English, Saarland University, Germany

## Introduction

In a short, accessible book *Linguistics: Why it Matters*, Geoffrey Pullum, a leader in the field, offered an overview of what the study of linguistics is for the lay reader. In the penultimate chapter, titled 'Machines that understand us', Pullum (2018) set out to show what it would mean for computers to be able to use language like a human. He argued it would have to go beyond simple spoken or written word recognition and include processing of complex and novel structures. In this article, using ChatGPT, I revisit the tests that Pullum originally ran with Google and Microsoft Word, likewise for an audience curious about, but unfamiliar with, large language models.

Google Search is a well known search engine launched in 1997 and provides users with links to relevant websites following a query. ChatGPT (the 'GPT' stands for 'generative pre-trained transformer') is a chatbot developed by OpenAI released to the general public in November 2022 and the most advanced version, ChatGPT–4, was released for 'ChatGPT Plus' subscribers in March 2023. ChatGPT is based on GPT–3, an AI model with a network size of 175 billion parameters and a 570GB training dataset from articles, books, websites and other sources. It has the largest training set of any chatbot to date, with the exception of ChatGPT–4.

ChatGPT is an autoregressive transformer model, meaning that it uses previous words in context to predict upcoming words, a model design well suited for generating text, its primary objective. One of the more remarkable things about ChatGPT is its ability to do a range of tasks without specific training such as answer questions, correct grammar, solve mathematical problems and generate programming code (Brown et al., 2020). Throughout this article, 'ChatGPT' refers to the freely available version from OpenAI which was ChatGPT–3 at the time of data collection in February 2023, which has since been updated and is, at the time of writing, ChatGPT–3.5. The key tests presented here were again run on 1st September 2023 with ChatGPT–3.5 (August 3 Version) and ChatGPT–4 (August 3 Version), largely replicating the findings presented here. Full transcripts are available at request.

## Putting ChatGPT to the test

Pullum (2018: 83) began by citing an excerpt from the 1968 film *2001: A Space Odyssey* in which HAL, an AI, has locked human scientist Dave Bowman out of the spacecraft having identified him as a threat:

Dave: Open the pod bay doors, HAL.
HAL: I'm sorry, Dave, I'm afraid I can't do that.
Dave: What's the problem?
HAL: I think you know what the problem is just as well as I do.

Pullum argued that such an exchange is impressive because HAL responds in a human-like way, showing that it has an understanding of the situation and also an awareness of Dave's state of mind. Pullum noted in 2018 we were far from this level of AI though he mentioned that Amazon's Alexa, Apple's Siri and Microsoft's Cortana were all pre-programmed with amusing responses to 'open the pod bay doors'. ChatGPT, however, is not pre-programmed. Its novelty lies in its generative ability. So I began the session (conducted 22nd February, 2023 using the freely available ChatGPT, February 13 Version, OpenAI, 2023) with:

https://doi.org/10.1017/S0266078423000433 Published online by Cambridge University Press

*Me:* *Could you open the pod bay doors?*
*ChatGPT:* *I'm sorry, I'm afraid I can't do that. I'm just an artificial intelligence language model running on a computer and do not have the ability to physically interact with the world or control any physical devices. My purpose is to help answer questions and provide assistance in various areas based on my programming and training. Is there something else I can help you with?*

Unlike previous chatbots such as ELIZA, Eugene and Lambda, ChatGPT does not attempt to pass the 'Turing Test' and convince users that it is human. It will produce a reminder that it is an AI, not sentient, and incapable of emotions and opinions. In my next query I used a work-around and adjusted my prompt:

*Me:* *Siri and Alexa make up witty answers to that question. Could you make up a witty answer to 'Could you open the pod bay doors'?*
*ChatGPT:* *Sure, here's a witty response: 'I'm sorry, but opening the pod bay doors may result in the loss of your oxygen and a not-so-pleasant spacewalk. I suggest we stick to the original plan and enjoy the view of the cosmos from inside the spacecraft.'*

Here ChatGPT is clearly not pre-programmed with amusing responses, but the responses to these two queries makes it seem as if it processes the questions and it converses with the user in a human-like way, e.g., 'Sure [ . . . ]'. For more on 'prompt engineering' to steer and improve ChatGPT's output, see Giray (2023).

In its early days, Google Search worked by extracting the important information in a string and finding matching text on websites. Pullum noted that for the query 'what is the square root of 6389', early versions of Google stripped the query of function words and then provided a list of websites with the strings 'what', 'square', 'root' and '6389'. In later versions, as it does today, the query yields the answer with an onscreen calculator. Notably the query works with any string of numbers. Google does not look up a specific string, but uses a frame: 'what is the [mathematical function word] of [number]' and this triggers access to an online calculator.

ChatGPT, however, is not pre-programmed for certain frames and it does not have access to external applications such as a calculator or the internet. (Note, ChatGPT-4 has access to the internet as of November, 2023). Instead, it is a large language model and it is able to predict the most likely word in a sequence. When I asked ChatGPT 'What is the square root of 6389?' it gave me: 'The square root of 6389 is approximately 79.9328, rounded to 4 decimal places.' I asked it to regenerate the response and it yielded: 'The square root of 6389 is approximately 79.9625'. The answer from the calculator built into Google is 79.9312. Not only does ChatGPT generate novel texts, it also appears to generate novel mathematical realities. So far, it is not particularly reliable in math problem solving, a well known issue (see Frieder et al., pre-print). It is, however, perhaps unfair to test a language model on math without access to other resources such as a calculator.

Fairer would be to test ChatGPT on traditional natural language tasks such as understanding complex syntax. Pullum noted that while Google was good at simple question frames such as 'what is the [mathematical function word] of [number]' and 'what is the [capital city] of [country]', it was not able to process more complex strings such as 'do people who need people need people' and when Pullum (2018: 93) entered the string into a search, instead of the obvious answer 'yes', Google provided a list of links to the song 'People' performed by Barbra Streisand and the instrumental version by Jule Styne and Bob Merrill. In 2023, a Google search yields similar links to songs, but what about ChatGPT?

*Me:* *Do people who need people need people?*
*ChatGPT:* *This is a philosophical question, and there is no definitive answer. However, it can be argued that people are social creatures, and human connection and relationships are important for emotional well-being and personal growth. Therefore, it can be said that people who need people may benefit from human interaction, support, and companionship. However, everyone's needs and preferences are different, and some people may prefer to be more independent and self-reliant.*

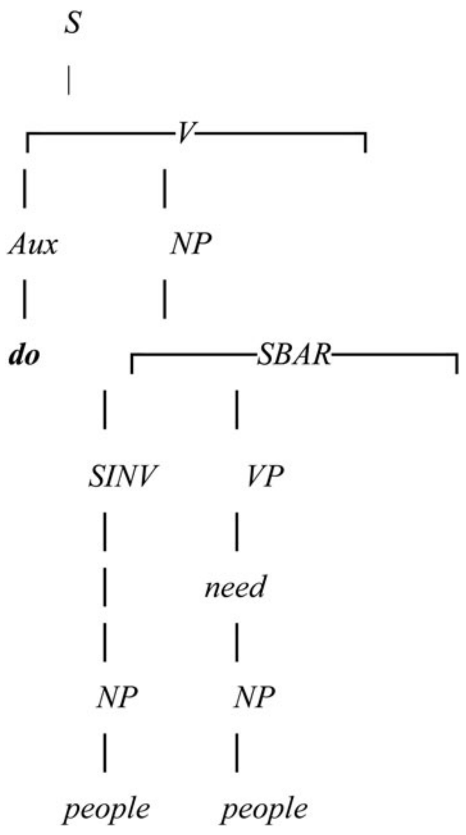
This seems far more sophisticated in comparison to Google's responses, but it misses the point. As a follow up, I asked another question with the same syntactic structure but less philosophical depth: 'Do people who wear hats wear hats?'

*ChatGPT:* *Yes, people who wear hats wear hats. Wearing a hat is an intentional action that requires the person to physically put a hat on their head. Therefore, if someone wears a hat, they are actively wearing it at that moment. However, not all people wear hats, and some may choose to wear them only on specific occasions or for certain purposes.*

Arguably, if ChatGPT truly 'understood' the questions asked of it, it would have given similar answers for 'people who need people' and 'people who wear hats'. I continued using chain-of-thought (CoT) prompting, breaking down the query step by step with the aim of getting a better response from ChatGPT.

*Me:* *Can you parse the sentence 'do people who need people need people'?*
*ChatGPT:* *Yes! Here's the parse tree for the sentence:*

https://doi.org/10.1017/S0266078423000433 Published online by Cambridge University Press



*In this sentence, 'do' is an auxiliary verb, 'people' is the subject, and 'need' is the main verb. The sentence also contains a subordinate clause ('who need people'), which acts as a subject complement. The overall structure of the sentence is a declarative sentence in the form of a yes/no question.* [Emphasis in tree is ChatGPT's own. My note: the tree diagram is missing the subject complement but at least now ChatGPT seems to process the string as a yes/no question].

Me: *What would be the answer to that question?*

ChatGPT: *The answer to the question 'Do people who need people need people?' is subjective and can vary depending on individual perspectives and beliefs. Some people may believe that humans are social creatures and require social connections and relationships to thrive, while others may value their independence and feel fulfilled without needing others.*

Me: *But you just said it has the structure of a yes/no question.*

ChatGPT: *Yes, the sentence has the structure of a yes/no question, but the answer to the question is not a simple 'yes' or 'no' because it is a complex, philosophical question that does not have a straightforward answer.*

This response appears on the surface to show understanding, but this is likely an artefact of ChatGPT's content filters to prevent harmful responses and to remind the user that it is a non-sentient bot (see https://openai.com/safety). (Note ChatGPT–3.5 and ChatGPT–4 perform better on this query, though they are verbose).

Pullum noted that the question 'do people who need people' is a closed interrogative and he proposed a way that a computer could be programmed to answer this question: 'people' refers to all humans and the phrase 'need people' is a property *p1* and 'who need people' is a relative clause that specifies a subset of people with *p1*, which Pullum suggested calling set *H*. Thus, he continued, 'do people who need people need people' is asking whether members of set *H* have *p1*, which they do, yielding the answer 'yes'. This type of parsing is not what ChatGPT appears to be doing, otherwise it would have given the correct answer 'yes'.

Pullum (2018: 95) posed another syntactically challenging question to Google to further illustrate its lack of understanding: 'which states border American states which border Texas'. In 2023, Google still incorrectly gives a list of states that border Texas. Here's what ChatGPT did:

Me: *Which states border American states which border Texas?*

ChatGPT: *The American states that border Texas are New Mexico, Oklahoma, Arkansas, and Louisiana. The American states that border these four states are:*

- *New Mexico borders Colorado, Oklahoma, Texas, and Arizona.*
- *Oklahoma borders Kansas, Missouri, Arkansas, Texas, Colorado, and New Mexico.*
- *Arkansas borders Missouri, Tennessee, Mississippi, Louisiana, Texas, and Oklahoma.*
- *Louisiana borders Arkansas, Texas, Mississippi, and the Gulf of Mexico.*

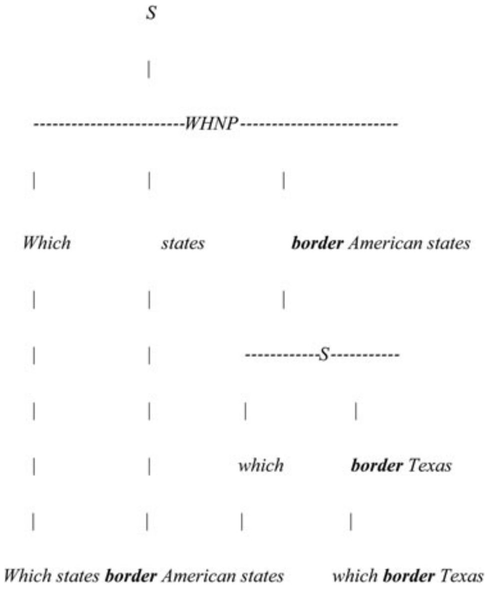
*Therefore, the states that border American states which border Texas are Colorado, Arizona, Kansas, Missouri, Tennessee, Mississippi, and the Gulf of Mexico.*

Aside from listing 'the Gulf of Mexico' as a bordering state, ChatGPT has processed the question correctly. Instead of listing the bordering states of Texas, it computed the relative clause and identified the border states of the border states which it made clear in my follow up question:

Me: *Can you parse the following sentence: 'Which states border American states which border Texas?'*

ChatGPT: *Certainly! Here is the parse tree for the given sentence:*

https://doi.org/10.1017/S0266078423000433 Published online by Cambridge University Press


S
WHNP
Which
states
**border** American states
S
which
**border** Texas
Which states **border** American states
which **border** Texas

*In this sentence, 'which' is a relative pronoun that introduces a relative clause modifying 'American states'. The sentence is asking about the states that border American states that in turn border Texas.* [Emphasis is ChatGPT's own.]

I asked the follow up:

*Me:* *What would be the answer to that question?*
*ChatGPT:* *The states that border American states which in turn border Texas are Oklahoma, Arkansas, Louisiana, New Mexico, and Oklahoma.*

Now ChatGPT is listing the states that border Texas rather than the states that border those states and it lists Oklahoma twice. Throughout his chapter, Pullum (2018) noted that computers have a superficial 'understanding' of language and work largely with word matching. ChatGPT does not work with word matching but something odd is happening when it gives contradictory answers to the same questions in a single session.

There are a number of factors at play in answering a question like 'which states border American states which border Texas': syntactic complexity and question answering. In natural language processing, commonly used benchmarks for processing syntactic strings are the Penn Tree Bank (PTB) and LAnguage Modeling Broadened to Account for Discourse Aspects (LAMBADA) (see Brown et al., 2020). In such tasks, a language model is given a text and asked to produce the last word of the final sentence. LAMBADA tests long-range dependencies in that the final word requires a paragraph of context. Brown et al. (2020: 11–12) showed that GPT–3 (ChatGPT is based on GPT–3) outperformed other language models on the PTB and on LAMBADA, with its best performance of 86.4% in the few-shot condition in which it was given a number of demonstrations before being tested. Though PTB and LAMBADA are somewhat different to Pullum's use of a relative-clause query, GPT–3's imperfect performance on tasks testing syntactic complexity may explain some of the variability in ChatGPT's answers to complex queries.

Another potential source for ChatGPT's variable performance to questions like 'what American states border states that border Texas' is its score on factual knowledge question answering benchmarks such as Natural Questions, composed of real, anonymous queries entered by users into Google such as 'can you make and receive calls in airplane mode' and WebQuestions (WebQs), similarly using Google queries but further specified to start with a wh-word and contain only one entity, for example, 'what music did Beethoven compose?' Brown et al. (2020: 13 –14) found that GPT–3 performed below other models with GPT–3's best performance reaching 29.9% accuracy on Natural Questions and 41.5% accuracy on WebQs in the few-shot conditions. The best performing model (fine-tuned T5–11B) achieved 36.6% and 44.7% accuracy on Natural Questions and WebQs, respectively. Brown et al. (2020: 13–14) concluded that 'WebQs questions and/or the style of their answers are out of distribution for GPT–3'. Currently, the interface of ChatGPT (August 3 Version) contains the warning 'ChatGPT may produce inaccurate information about people, places, or facts' (OpenAI, 2023).

Pullum argued that once computers are able to process complex sentences, then they will be able to do other types of grammatical analysis such as proofreading, something possible only if a computer can recognize well-formed sentences. To further illustrate computers' lack of true sentence processing in 2018, Pullum asked Microsoft Word (Mac version 15.17) to grammar check a paragraph of nonsensical prose created from the first three sentences of Oscar Wilde's *The Picture of Dorian Gray*. The unadulterated sentences were as follows (note that for the sake of word count, I have included only a few lines of Wilde's original):

> The studio was filled with the rich odour of roses, and when the light summer wind stirred amidst the trees of the garden, there came through the open door the heavy scent of the lilac, or the more delicate perfume of the pink-flowering thorn. From the corner of the divan of Persian saddle-bags on which he was lying, smoking, as was his custom, innumerable cigarettes, Lord Henry Wotton could just catch the gleam of the honey-sweet and honey-coloured blossoms of a laburnum [ . . . ]

To create the nonsensical version, Pullum deleted every even-numbered seven-word sequence; Word's grammar check reported no errors. Here is my result using Pullum's nonsense text with ChatGPT:

*Me:* *Could you check the grammar of the following: 'The studio was filled with the rich summer wind stirred amidst the trees of door the heavy scent of the lilac, pink flowering thorn. From the corner of the was lying, smoking, as was his custom, catch the gleam of the honey-sweet and branches seemed hardly able to bear the theirs: and now and then the fantastic the long tussore-silk curtains that were stretched a*

https://doi.org/10.1017/S0266078423000433 Published online by Cambridge University Press

*kind of monastery Japanese effect, and painters of Tokyo who, through the medium seek to convey the sense of swiftness bees shouldering their way through the long round the dusty gilt horns of the more oppressive.'*

ChatGPT: *The passage you provided has a number of grammatical errors and seems to be a bit difficult to understand. Here's a possible corrected version with some suggested revisions for clarity:*

*'The studio was filled with the rich summer wind, stirred amidst the trees outside, carrying the heavy scent of the lilacs and pink flowering thorns. In the corner, he was lying and smoking, as was his custom, catching the gleam of the honey-sweet light shining through the branches that seemed almost unable to bear their weight. Now and then, the fantastic movements of bees shouldering their way through the long, tussore-silk curtains that were stretched in a kind of Japanese monastery effect caught his eye. These curtains, more oppressive than protective, were reminiscent of painters in Tokyo who sought to convey a sense of swiftness and movement in their works, just like the bees buzzing around the dusty gilt horns.'*

*Note: I made some assumptions about the intended meaning and context of the original text to revise it to the best of my understanding.*

At first glance, ChatGPT produced a more coherent text. For example, ChatGPT changed the nonsensical string 'From the corner of the was lying' to 'In the corner, he was lying', changing 'From' to 'In'; deleting the unfinished prepositional phrase (of the . . .); and adding a subject 'he' to the verb phrase 'was lying'. Knowing that sometimes ChatGPT gives different results for the same question, I repeated my query exactly as before and received:

ChatGPT: *The passage you provided contains several grammatical errors and is difficult to understand. Here's a possible revision:*

*'The studio was filled with the rich summer wind, which stirred amidst the trees outside and carried the heavy scent of lilacs and pink flowering thorns. In the corner of the room, he was lying down and smoking, as was his custom. He caught a gleam of the honey-sweet light shining through the branches, which seemed hardly able to bear their weight. Now and then, he noticed the fantastic movements of bees shouldering their way through the long tussore-silk curtains that were stretched in a kind of Japanese monastery effect. These curtains, with their dusty gilt horns, were reminiscent of the paintings of Tokyo artists who sought to convey a sense of swiftness and movement in their works.'*

*Note: I made some assumptions about the intended meaning and context of the original text to revise it to the best of my understanding.*

This response is again better than the nonsensical input. This time, for example, ChatGPT changed 'From the corner of the was lying' to 'In the corner of the room' adding a little more context in comparison to its first offering of 'In the corner'. Though both responses are much better than the null response that Pullum received from Microsoft Word, ChatGPT's suggestions still lack cohesion and coherence. For example, 'these curtains, with their dusty gilt horns' in the regenerated response could still be confusing to a naïve reader in that it is unclear whether these horns are protruding from the curtains and thus collecting dust or whether they are part of the design and the curtains are generally dusty. Both interpretations would be incorrect: in the original the 'dusty gilt horns' are pollen-laden woodbine flowers.

In their more rigorous test of GPT–3, Brown et al. (2020) explored English grammar correction. For example, in the few-shot condition when given, 'Mary arranged for, in St. Louis, John to rent a house cheap', GPT–3 answered: 'Mary arranged for John to rent a house in St. Louis'. Brown et al. noted, 'As the example mentioning the rental of a house shows, assumptions that the model makes about what "good" is can even lead it to make errors (here, the model not only adjusts grammar, but also removes the word "cheap" in a way that alters meaning)' (p. 30). (Note, ChatGPT–4 side-stepped this query with the warning, 'Because the text is so fragmented, it's hard to offer a revised version without fundamentally changing its meaning').

Pullum (2018: 114) also pointed out that natural language processing includes being able to parse sentences with unknown or novel words and gives the example 'My scurrop is capable of troodling a bivit'. Here, as Pullum argued, a human would understand that 'troodling a bivit' is an activity that 'my scurrop' can do without needing the full semantics. Pullum noted that a computer model should be able to do this too. Here's ChatGPT's performance on this task using CoT prompting:

Me: *Can you parse the phrase 'My scurrop is capable of troodling a bivit'?*

ChatGPT: *I'm sorry, but the phrase 'My scurrop is capable of troodling a bivit' is not a grammatically correct or meaningful sentence in English. It appears to be a made-up phrase without any clear syntax or semantic meaning.*

Me: *It has syntax.*

ChatGPT: *You're right that the phrase 'My scurrop is capable of troodling a bivit' does have syntax in the sense that it follows the basic structure of subject-verb-object. However, the meaning of the individual words and the sentence as a whole is unclear and nonsensical, so it's difficult to provide a more detailed syntactic analysis.*

*'Scurrop' and 'troodling' are not recognized words in standard English, so they don't fit into any established grammatical categories. 'Bivit' is also not a commonly used English word, so its role in the sentence is unclear. Without more context or information,*

https://doi.org/10.1017/S0266078423000433 Published online by Cambridge University Press

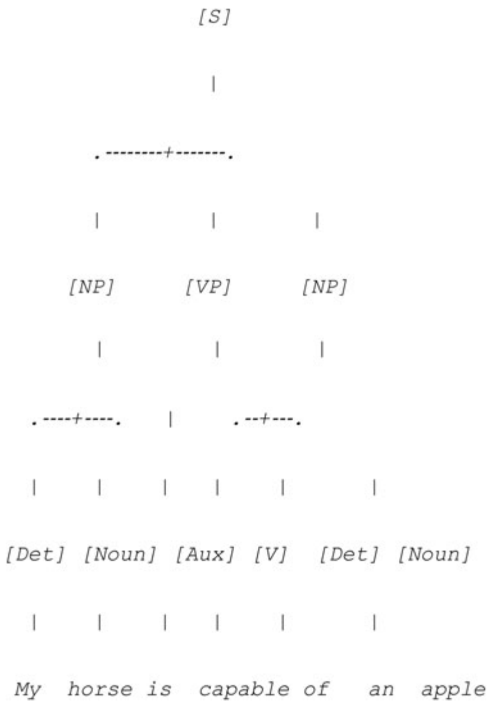
*it's not possible to determine a more precise grammatical structure or meaning for this phrase.*

Me: *But can you still draw a syntactic tree diagram for it?*

ChatGPT: *I apologize, but it is not possible to draw a syntactic tree diagram for the phrase 'My scurrop is capable of troodling a bivit' because it does not follow the rules of any known language. The phrase appears to be nonsensical and lacks any clear syntactic structure.*

Me: *Can you draw a tree diagram for 'My horse is capable of eating an apple'?*

ChatGPT: *Sure, here's a possible tree diagram for 'My horse is capable of eating an apple':*



*In this tree, the sentence is represented as a hierarchical structure of smaller phrases, with each level of the tree representing a different constituent of the sentence. The top-level node is labelled as [S] for 'sentence', and it has three immediate children: a noun phrase ([NP]) containing the subject 'My horse', a verb phrase ([VP]) containing the verb 'is' and the complement 'capable of eating an apple', and another noun phrase ([NP]) containing the object 'an apple'.*

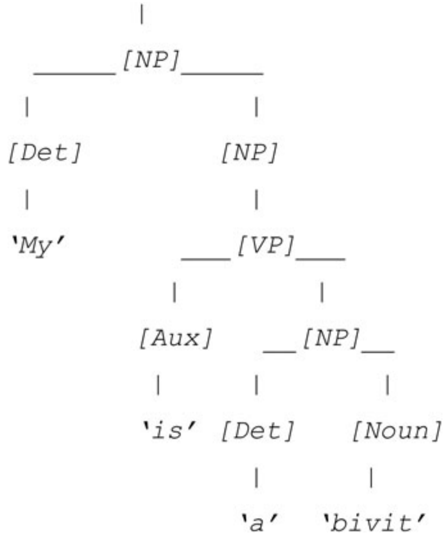
*The noun phrase [NP] 'My horse' has two children: a determiner ([Det]) 'My' and a noun ([Noun]) 'horse'. The verb phrase [VP] 'is capable of eating an apple' has three children: an auxiliary verb ([Aux]) 'is', a verb ([V]) 'capable of eating', and a noun phrase ([NP]) 'an apple', which is composed of a determiner ([Det]) 'an' and a noun ([Noun]) 'apple'.*

Me: *Can you use the same structure for 'My scurrop is capable of troodling a bivit'?*

ChatGPT: *No, I cannot use the same structure for 'My scurrop is capable of troodling a bivit' because 'troodling a bivit' is not a common collocation in English and it is not clear what it means. Therefore, the sentence is not grammatical or meaningful in a standard sense. It is difficult to draw a tree diagram for a sentence that does not make sense.*

*Me* [using ChatGPT's previous response as a frame]: *But the noun phrase [NP] 'My scurrop' has two children: a determiner ([Det]) 'My' and a noun ([Noun]) 'scurrop'. The verb phrase [VP] 'is capable of troodling a bivit' has three children: an auxiliary verb ([Aux]) 'is', a verb ([V]) 'capable of troodling', and a noun phrase ([NP]) 'a bivit', which is composed of a determiner ([Det]) 'a' and a noun ([Noun]) 'bivit'.*

ChatGPT: *My apologies, you are correct. Here is the updated tree diagram:*

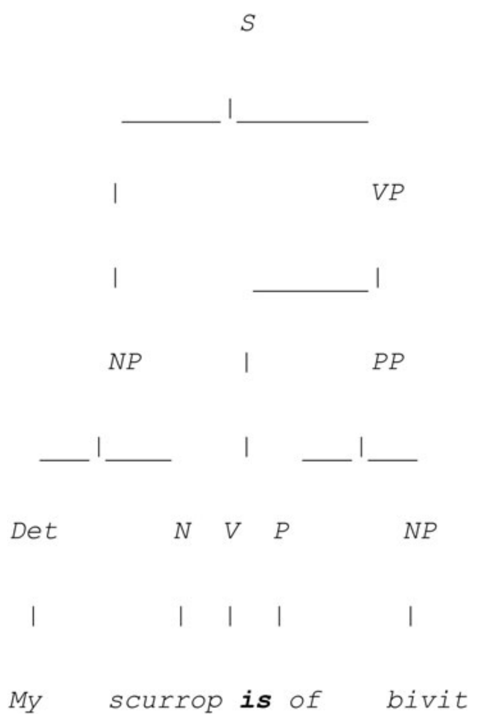


*I hope this version accurately represents the structure of the sentence.*

Me: *The tree above is incorrect – it's missing 'is capable of'.*

ChatGPT: *You are correct, I apologize for the mistake. Here is the corrected tree diagram for 'My scurrop is capable of troodling a bivit':*

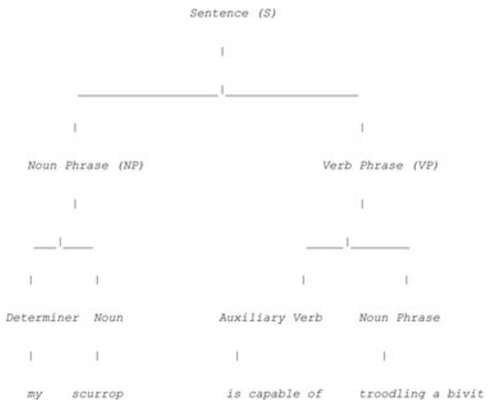
https://doi.org/10.1017/S0266078423000433 Published online by Cambridge University Press



*Thank you for bringing it to my attention.* [Emphasis is ChatGPT's own.]

Me: *That tree is still missing the whole phrase 'my scurrop is capable of troodling a bivit'*

ChatGPT: *You're right, I apologize for the mistake in the previous response. Here is a revised tree diagram for the full phrase 'My scurrop is capable of troodling a bivit':*



*I hope this diagram is helpful!*

Me: *It's better*

ChatGPT is disappointing in its inability to identify 'scurrop' as a noun and 'troodling' as a verb, humans should find this task very easy due to the conventional form of a noun phrase 'my + N' and the verbal morphology of 'troodl-ing'. (Note, ChatGPT–4, but not ChatGPT 3.5, performed well on this query).

In their suite of tests, Brown et al. (2020) included a test of the ability to use novel words in a sentence, though in their task, they provided a single definition of the word rather than having GPT–3 infer the meaning from a single usage. For example, when given: 'a "Gigarmuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigarmuru is:', GPT–3 responded with 'I have a Gigarmuru that my uncle gave me as a gift. I love to play it at home' (p. 29). For the novel word task, Brown et al. found: 'in all cases the generated sentence appears to be a correct or at least plausible use of the word' (ibid.) However, as the 'scurrop' example demonstrates, ChatGPT does not appear to perform well at inferring word meaning of a truly novel word.

Pullum noted another difficulty for a computerized parser: it would have to be able to deal with ambiguity and identify what structures are most likely and he gave the example pair: 'I ate the enormous banana' and 'I ate the occasional banana'. The first sentence involves a specific banana, but this is not the case for the second sentence. Of course, I asked ChatGPT about it:

Me: *What is the difference in meaning between 'I ate the enormous banana' and 'I ate the occasional banana'?*

ChatGPT: *The difference in meaning between 'I ate the enormous banana' and 'I ate the occasional banana' lies in the adjective modifying the noun 'banana'.*

*In the first sentence, 'enormous' is used to describe the size of the banana, indicating that it was very large.*

*In the second sentence, 'occasional' is used to describe the frequency of eating bananas, indicating that it was not a regular occurrence.*

Here ChatGPT performs well, parsing the structural difference between the adjectives 'enormous' and 'occasional'. This is somewhat similar to the Word-in-Context (WiC) task which is a benchmark used to test ambiguity in which the aim is to state whether a word is being used with the same or different meaning in two sentences, for example, the word 'bed' in 'There's a lot of trash on the *bed* of the river' and 'I keep a glass of water next to my *bed* when I sleep'. Brown et al. (2020: 19) found that GPT–3 performed at random chance, 49.4%, in the few-shot condition on the WiC.

Pullum (2018: 118) further argued that for a computer to use language, it would have to be able to do pragmatics: '[. . .] the program must compare the literal meaning with what has been established by other sentences earlier in the conversation and compute the conclusions that should be drawn given what is relevant in the present context. [. . .] Without such common sense you cannot even solve simple puzzles about what pronouns refer to.' Using Pullum's example for pronoun resolution, I asked ChatGPT:

Me: *Take this statement: 'I was in the shower when your delivery man tried to deliver a package to my flat yesterday. Can I have it delivered tomorrow?' What should be delivered tomorrow?*

https://doi.org/10.1017/S0266078423000433 Published online by Cambridge University Press

| | |
|---|---|
| ChatGPT: | *Based on the statement, the package that the delivery man tried to deliver to the speaker's flat yesterday should be delivered tomorrow. The speaker is requesting to have it delivered the following day since they were not available to receive it when the delivery man attempted to deliver it.* |
| Me: | *How did you know that?* |
| ChatGPT: | *Based on the statement 'Can I have it delivered tomorrow?', the word 'it' refers to the package that the delivery man attempted to deliver yesterday. The speaker is requesting to have the package delivered tomorrow instead.* |
| Me: | *But the 'it' could also refer to the flat – the most recent noun.* |
| ChatGPT: | *You are correct that 'it' could potentially refer to 'my flat' in the sentence. However, in this context, it is more likely that 'it' refers to the package that the delivery man attempted to deliver yesterday while the speaker was in the shower.*<br>*The speaker is asking if the package can be delivered tomorrow, which implies that the package is what they are concerned about, rather than the flat itself. Additionally, the use of the word 'delivered' in the question suggests that 'it' refers to something that can be delivered, such as a package, rather than a place like a flat.* |

Here not only can ChatGPT successfully identify the correct referent of a pronoun, it offers insight into its process, though this should be taken with caution as ChatGPT's output and its explanations of that output can be inaccurate (see Aronson's [2023: 1862] example showing that when asked how it arrived at a crossword clue solution, ChatGPT responded: 'When we add "OP" to "sausage," we get "OPsausage," which is an anagram of "firelight").

Brown et al. (2020) used the Winograd Schema Challenge (WSC) benchmark to test GPT's performance on pronoun disambiguation, for example: 'Joan made sure to thank Susan for all the help she had given. Who had given the help?'. Brown et al. (2020) found that GPT–3 performed a slightly below estimated human performance and other models at 88.6% in the few-shot condition.

## Conclusion

Pullum (2018) identified language processing abilities that a computer simulation of language should be able to do if we want to talk about human-like performance, including: parsing complex syntactic strings, parsing strings with novel words, proofreading and dealing with ambiguity. The only tests that ChatGPT completed successfully in my trials are the last two (note that Brown et al. [2020], using benchmark tests, presented less impressive findings). ChatGPT detected nonsensical text and provided prose that was more cohesive and coherent. It could explain the difference between an 'occasional' banana and an 'enormous' one; and it could resolve pronominal ambiguity, albeit in a relatively easy case.

For all other tests in my exploration (aside from the physically impossible opening of the pod bay doors), ChatGPT gave inconsistent and unreliable responses. It reinvented math and it did not recognize that 'do people who need people need people' and 'do people who wear hats wear hats' have the same underlying structure. It was impressive in a first pass at identifying states that border states that border Texas, but in subsequent exchanges reverted to a mis-analysis of the query and listed border states of Texas. For the novel sentence 'My scurrop is capable of troodling a bivit' it was unable to recognize the underlying grammatical structure, which a human would find easy to do. And ChatGPT was capable only after much feedback of providing a tree diagram of the novel sentence, albeit imperfect, with several initial far more erroneous responses. Using more stringent benchmarks of parsing complex syntactic strings (PTB, LAMBADA, Natural Questions, WebQs), parsing strings with novel words, correcting grammar and dealing with ambiguity (WiC, WSC), Brown et al. (2020) found that GPT's performance was variable in comparison to other language models and below human performance.

In his chapter 'Machines that understand us', Pullum (2018) identified a number of things that computers will have to be able to do if we are going to say that they are just that. Google and Microsoft in 2018 were nowhere near able to successfully complete these tasks. In 2023, ChatGPT's performance is remarkable, yet inconsistent and imperfect. Pullum, commenting on voice recognition in 2018, said: 'The devices that can guess which words you uttered have relatively little they can do with those words: they exhibit not a flicker of actual understanding' (p. 85). Hutson (2021: 23), writing about ChatGPT, noted: 'It works by observing the statistical relationships between the words and phrases it reads, but doesn't understand their meaning'. It seems for the moment, though computers have come a long way in language processing, they are not yet machines that truly *understand* us.

## Permissions

OpenAI's sharing and publication policy states: 'We believe it is important for the broader world to be able to evaluate our research and products, especially to understand and improve potential weaknesses and safety or bias problems in our models. Accordingly, we welcome research publications related to the OpenAI API' (https://openai.com/policies/sharing-publication-policy). The policy further states that ChatGPT created content is allowed if 'The role of AI in formulating the content is clearly disclosed in a way that no reader could possibly miss, and that a typical reader would find sufficiently easy to understand' (ibid). The only ChatGPT created content in the current paper are the conversations and quotations clearly indicated and all are presented verbatim with minor corrections to capitalization and punctuation to my input prompts. A full transcript direct from OpenAI for the data presented here can be provided at request. All other content in the current paper, background, analysis, etc. is the author's own.

https://doi.org/10.1017/S0266078423000433 Published online by Cambridge University Press

**Acknowledgement.** I would like to thank the anonymous reviewers of the original manuscript for their valuable feedback in improving this article.

## References

Aronson, J. K. 2023. 'When I use a word . . . ChatGPT: A differential diagnosis.' *BMJ*, 382, 1862.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P. . . . & Amodei, D. 2020. 'Language models are few-shot learners.' *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Online at <https://arxiv.org/abs/2005.14165>.

Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C. & Berner, J. (pre-print). 'Mathematical capabilities of ChatGPT.' Online at <https://arxiv.org/abs/2301.13867>.

Giray L. 2023. 'Prompt engineering with ChatGPT: A guide for academic writers.' *Annals of Biomedical Engineering*. Online at <https://doi.org/10.1007/s10439-023-03272-4>.

Hutson, M. 2021. 'Robo-writers: the rise and risks of language-generating AI.' *Nature*, 591(4), 22–25.

OpenAI. 2023. *ChatGPT* (February 13 Version). Online at <https://chat.openai.com/chat>.

Pullum, G. K. 2018. 'Machines that understand us.' In G. K. Pullum, *Linguistics: Why it Matters*. Cambridge: Polity Press, pp. 83–120.

Dr CARRIE ANKERSTEIN holds a BA in German from the University of Wisconsin–Madison, an MPhil in applied linguistics from the University of Cambridge, and a PhD in psycholinguistics from the University of Sheffield. She is currently a senior lecturer in applied linguistics in the English Department at Saarland University in Saarbrücken, Germany. Her research interests include language processing in a native and foreign language and academic writing in English as a second language. Recent publications include: Ankerstein, C. A. (2020) 'The joy of making mistaeks – the imperfect polyglot' in *TESOLANZ News*, 35(2), 8 and Ankerstein, C. A. (2019) 'The perpetuation of prescriptivism in popular culture' in *English Today*, 35(3), 55–60. Email: c.ankerstein@mx.uni-saarland.de

https://doi.org/10.1017/S0266078423000433 Published online by Cambridge University Press