ChatGPT Is Not Just ‘Good Enough’ But ‘Best Enough’: Comparing Garden-Path Processing in GPT-3.5, GPT-4, and Humans.

This study explores the syntactic processing of Large Language Models (LLMs) such as GPT-3.5 and GPT-4 and compare them with human processors, focusing on the garden-path (GP) sentences. While previous studies have shown that models such as BERT can handle GP with human-like syntactic representation, there has been few known studies specifically comparing GPT-3.5 and GPT-4 in this context. The intrigue of GP lies in the challenge they pose even to proficient human processors, often causing lingering misinterpretations despite subsequent reanalysis [1] [2]. This showcases the incomplete or ‘good-enough’ nature of human syntactic processing. This research seeks to determine whether LLMs exhibit a similar ‘good-enough’ syntactic processing as humans, or if their distinct, input-reliant learning mechanisms lead to more accurate or non-syntactic driven interpretations.

To this end, this study applied a comprehension questions task (task 1) and a paraphrasing task (task 2) to GPT models, following [1] and [2]. In Task 1, the models were presented with sentences such as “(a) While the man hunted the deer ran through the woods,” and responded ‘yes’ or ‘no’ to questions like “(b) Did the man hunt the deer?” or “(c) Did the deer run through the woods?”. A ‘yes’ response to (b) would indicate an error due to initial misinterpretation, while a ‘yes’ to (c) would suggest a correct understanding of the main clause. A ‘yes’ to both questions suggests a coexistence of initial misinterpretation and correct reanalysis, similar to the ‘good enough’ processing. In Task 2, models were asked to paraphrase the sentences such as (a), which requires them to consider both subordinate and main clauses for interpretation. The models’ responses were scored as ‘full’ for accurate interpretations, ‘partial’ for coexisting misinterpretation and correct reanalysis, and ‘failed’ for the misinterpretations without reanalysis. The materials were manipulated for plausibility, length, head position and verb type to examine the models’ parsing from various aspects (Table 1, 2, 3). Both tasks used the latest versions of gpt-3.5 and gpt-4, executed via OpenAI’s API. Each sentence was queried ten times, with all questions posed as zero-shot prompts. Human response data was derived from [1] and [2].

In the results, (i) GPT models displayed both initial misinterpretation and proper reanalysis. In Task 1, all three sets of GP elicited significantly more errors compared to NGP sentences ($p<0.05$). However, the high accuracy in matrix questions across all sets indicates successful reanalysis, and in Task 2, the models produced a significantly higher number of partial analyses under GP than NGP ($p<0.5$). This suggests that the syntactic processing of GPTs aligns with the ‘good-enough’ characteristic of human syntactic processing. (ii) Further, as with humans, the models were influenced by plausibility, GP length, head position, and verb type ($p<0.05$). Misinterpretations increased with higher plausibility and longer GP regions, indicating a complex interplay with the structural elements of the sentences. Additionally, the influence of verb type, particularly the reduced misinterpretations with RAT verbs, underlines the models’ sensitivity to linguistic nuances. (iii) Comparing GPT-3.5 and GPT-4, there was little difference in handling GP structures; both models showed similar performance across all three sets in Task 1. However, in NGP conditions, GPT-4 exhibited a lower error rate compared to GPT-3.5 ($p<0.05$), and it also generated significantly more accurate full responses in Task 2 ($p<0.05$). This suggests that while more advanced models may not drastically differ in their ‘good-enough’ processing, they do demonstrate overall improved syntactic handling. (iv) Lastly, when compared with human performance, GPT-4 showed higher error rates in GPs, indicating a greater vulnerability to initial misinterpretation. However, its higher accuracy in matrix questions suggests that this is not solely due to syntactic processing failure. Interestingly, the trend reverses in paraphrase tasks where GPT-4 outperforms humans in producing full responses. This variability in their performance across tasks highlights their adaptable syntactic representation and suggests that their processing may embody even more ‘good-enough’ quality than human language processing.

Table 1. Material examples for Task 1 Set 1 (Plausibility & Length) - 42 items

Sentence
a. While the man hunted the deer (that was brown and graceful) ran into the woods. [GP – Plausible]
b. While the man hunted the deer (that was brown and graceful) paced in the zoo. [GP – Implausible]
c. While the man hunted the pheasant the deer (that was brown and graceful) ran into the woods. [NGP – Plausible]
Question: Did the man hunt the deer?

* GP= Garden-path, NGP= No Garden-path, inclusion of the word in parentheses indicates long conditions.

Figure 1. Bar graphs for Task 1 Set 1.

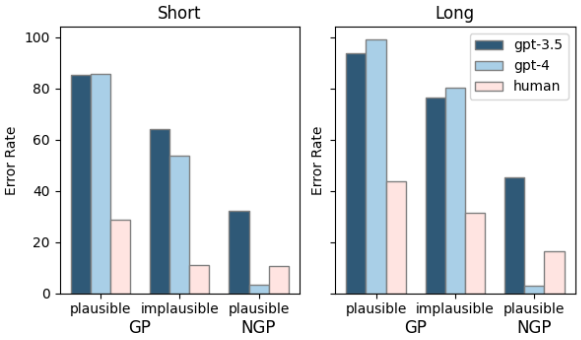


Table 2. Material examples for Task 1 Set 2 (Head & Main clause probing) - 40 items

Sentence
a. While the man hunted the deer that was brown and graceful ran into the woods. [GP – Head Early]
b. While the man hunted the brown and graceful deer ran into the woods. [GP – Head Late]
c. The deer that was brown and graceful ran into the woods while the man hunted. [NGP – Head Early]
d. The brown and graceful deer ran into the woods while the man hunted. [NGP – Head Late]
Subordinate Question : Did the man hunt the deer? Matrix Question : Did the deer run into the woods?

* In each trial, the model was presented with either a Subordinate or a Matrix Question, which resulted in a total set of 8 conditions.

Figure 2. Bar graphs for Task 1 Set 2.

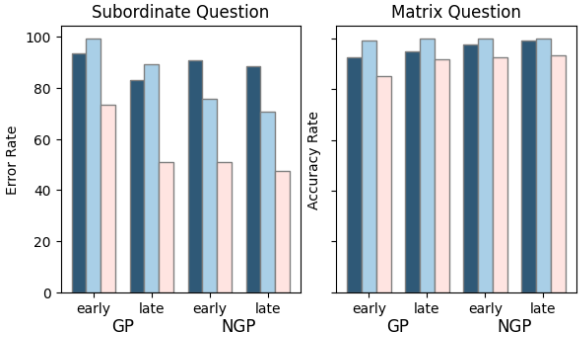


Table 3. Material examples for Task 1 Set 3 (Verb Type) and Task 2 - 12 items each for OT and RAT verb

Sentence
a. While the man hunted the deer that was brown and graceful ran into the woods. [GP – OT verb]
b. While the man hunted, the deer that was brown and graceful ran into the woods. [NGP – OT verb]
c. While Jim bathed the child that was blond and pudgy giggled with delight. [GP – RAT verb]
d. While Jim bathed, the child that was blond and pudgy giggled with delight. [NGP – RAT verb]
Question: Did the man hunt the deer? / Did Jim bathe the child? ('Yes' is incorrect response)

* OT= Optionally Transitive verb, RAT= Reflexive Absolute Transitive verb

Figure 3. Bar graphs for Task 1 Set 3.

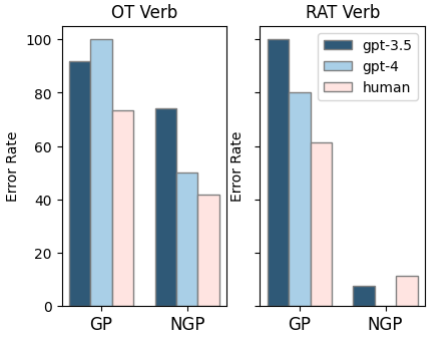
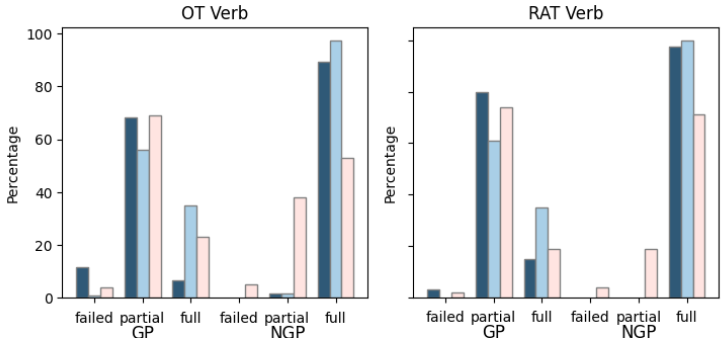


Figure 4. Bar graphs for Task 2.



[1] Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive psychology, 42*(4), 368-407.

[2] Patson, N. D., Darowski, E. S., Moon, N., & Ferreira, F. (2009). Lingering misinterpretations in garden-path sentences: evidence from a paraphrasing task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(1), 280.