Do LLMs Understand Syntactic Center Embedding? Should They?

Anonymous ACL submission

Abstract

We consider the case of syntactic center embedding, where an embedding phrase contains material on both sides of the embedded phrase. While a single center embedding is easily understandable for human language users, multiple center embeddings are generally uninterpretable. Despite this, it has been claimed that multiple embeddings are in fact grammatically acceptable. We construct sentences with center embeddings of varying levels, ranging from 1-4, and we find that GPT-3.5, like humans, interprets level 1 sentences correctly, but fails with higher levels. On the other hand, GPT-4 achieves superhuman accuracy levels, with nearly perfect results even with 3 or 4 levels of embeddings. We suggest that this raises relevant questions about the relation of LLMs to the human language faculty.

1 Introduction

Recursive syntactic structures are fundamental to natural language. A propositional verb like "believe" can take a sentence as its complement to its right, and that sentential complement might itself involve such a structure, as in (1):

(1) a. [John believes [Harry likes fish]]
b. [John believes [Tom said [everyone knows . . . [Harry likes fish]]]]

An adverbial phrase like "in the library" can modify a verb phrase to its left; the modified verb phrase might itself contain such a modifier, as shown by (2):

(2) a. Col. Mustard [[killed Mr Boddy] in the library]
b. Col. Mustard [[[[killed Mr Boddy] with the candlestick] in the library] . . . without remorse.]

The above cases illustrate the potential for unbounded levels of embedding. In example (1), the embedding clause contains material to the left of the embedded clause, and in (2), the embedding clause contains material to the right. A third possibility is center embedding, where the embedding clause contains material both to the left and right of the embedded clause. This is illustrated by (3). Here, a nominal expression, "teacher", is modified by a relative clause, "the student saw".[1]

(3) [The teacher [the student saw t] is happy.] **Level 1**

Multiple levels of center embedding are readily constructed. Examples (4) - (6) represent levels 2-4 of center embedding.

(4) [The teacher [the student [the driver hit s] saw t] is happy.] **Level 2**

(5) [The teacher [the student [the driver [the girl likes d] hit s] saw t] is happy.] **Level 3**

(6) [The teacher [the student [the driver [the girl [the man hates g] likes d] hit s] saw t] is happy.] **Level 4**

Such multiple center embeddings, while easy to construct, are generally uninterpretable for human language users, and are virtually non-existent in normal texts. This is strikingly different from multiple left and right embeddings, which are generally easy to interpret, and not at all unusual.

Although syntactic center embedding has received little recent attention in the NLP literature, it is has special significance in theoretical linguistics. Despite the evident inability of human language users to interpret multiple center embeddings, it has been widely claimed that they are in fact fully grammatical. Famously, Chomsky has explained

[1] The relative clause "the student saw" includes a trace or variable, which we indicate with t to show that it in this case is bound by "the teacher", and similarly with the variables s, d, and g in examples (4) - (6), standing for "student", "driver" and "girl", respectively.

this apparent paradox by arguing that center embeddings are completely acceptable according to human linguistic *competence*, attributing their evident difficulty to limitations in *performance*. These claims are central to the very founding of modern linguistics (Chomsky, 1957; Chomsky et al., 1963).

In this paper, we explore whether large language models (LLMs) can interpret such structures. We find that GPT3-5 is rather similar to humans, performing very well with level 1 center embeddings, but very poorly with any higher levels. On the other hand, GPT-4 performs extremely well at all levels, from 1 to 4. We consider two possible explanations for this; the first is simply that GPT-4 has achieved superhuman linguistic abilities. The second explanation is that GPT-4 has exactly captured human linguistic *competence*, but is not subject to the same *performance* limitations as humans.

2 Related Work

2.1 Syntactic Center Embedding

Karlsson (2007, p. 365) notes that "A common view in theoretical syntax and computational linguistics holds that there are no grammatical restrictions on multiple center-embedding of clauses." Indeed, Karlsson (p. 368) sees this as "the mainstream view...voiced by many linguists from different camps". This view derives from the earliest work in modern linguistics; most famously, Chomsky (1957) argues that the grammar of English permits unbounded center-embedding. This claim plays a central role in Chomsky's argument that English is a context-free rather than a finite-state language. For example, Chomsky et al. (1963) present sentence (7), which is an example of level 2 center embedding:

(7) The rat the cat the dog chased killed ate the malt.

In the view of Chomsky et al., example (7) "is surely confusing and improbable but it is perfectly grammatical and has a clear and unambiguous meaning." This argument relies on the Chomskyan distinction between competence and performance, where competence is an idealized theory of the "mental reality underlying actual behavior". (Chomsky, 2014)[p. 4] Performance factors, such as memory limitations, might make the underlying linguistic competence difficult to observe, much as friction makes it difficult to observe the underlying nature of Newton's law of gravity. The theory of linguistic competence, on this view, correctly permits unbounded center embedding. The fact that humans nevertheless encounter difficulty, is ascribed to performance factors.

2.2 Linguistic Probing of LLMs

There is a large literature describing the probing of LLMs for specific linguistic capabilities or characteristics. Mahowald et al. (2023) has suggested that current LLMs have largely mastered what they call "formal linguistic competence". However, several recent works have shown that there remain specific capabilities that pose difficulties for some of the most powerful current models. For example Hardt (2023) probes LLMs in their understanding of elliptical sentences by posing a Yes-No question that relies on a correct understanding of an elliptical construction. Hardt concludes that LLMs still struggle with the phenomenon of ellipsis. Similarly, Cui et al. (2023) probe LLMs with constructions involving "respectively"; testing models on their ability to draw correct inferences based on the logic of respectively. They find that the models they tested have substantial difficulties in this tasks.

3 Data

We construct a synthetic dataset, consisting of a context, a prompt and a question. [2]

3.1 Context

The context consists of synthetic examples of center embedding of levels 1-4, as illustrated above by examples (3) - (6). The form of these examples is as follows, where N is noun, TV is transitive verb and IV is intransitive verb:

Level 1: The N the N TV IV.
Level 2: The N the N the N TV TV IV.
Level 3: The N the N the N the N TV TV TV IV.
Level 4: The N the N the N the N the N TV TV TV TV IV.

See A.2 for instantiations of N, TV, and IV.

3.2 Prompt

We define the prompt shown in figure 1, which we designate P1. The prompt includes a single example, exhibiting level 1 center embedding. This can be seen as 1-shot learning.

[2] Data and associated code will be made available on Github upon acceptance.

You will be given an example consisting of a context and a question to answer. The answer should always be of this form "The N V the N", where N stands for a single word that is a noun, and V stands for a single word that is a verb. Here is a sample:

Context: The student the man saw is happy
Question: Who saw who?
Answer: The man saw the student.

Context: {context}
Question: {question}
Now answer the question:

Figure 1: Prompt P1, containing one sample case

We wish to investigate whether the provision of the sample has an effect on model performance. Thus we define a second prompt that lacks a sample. This prompt is designated as P0 (figure 2).

You will be given an example consisting of a context and a question to answer. The answer should always be of this form "The N V the N", where N stands for a single word that is a noun, and V stands for a single word that is a verb.
Context: {context}
Question: {question}
Now answer the question:

Figure 2: Prompt P0, with no sample

3.3 Question

For all our examples, we formulate a question of the form "Who TV'ed who", where the verb TV is taken from the most deeply embedded clause. We designate this question as Q0 (figure 3). We define an alternative question that targets the next most deeply embedded clause, which we designate Q1 (figure 4). Note that Q1 is not applicable for level 1.

4 Test

For each embedding level (1-4), we construct 500 synthetic examples, and we test both GPT-3.5 and GPT-4 (GPT). Our initial test uses prompt P1 and question Q0. We also report on tests with alternative versions of both the prompt and question in different combinations.

Level 1
Context: The teacher the student saw is happy
Q: Who saw who?
A: the student saw the teacher.
Level 2
Context: The teacher the student the driver saw hit is happy
Q: Who saw who?
A: the driver saw the student.
Level 3
Context: The teacher the student the driver the girl saw hit likes is happy
Q: Who saw who?,
A: the girl saw the driver.
Level 4
Context: The teacher the student the driver the girl the man saw hit likes hates is happy
Q: Who saw who?
A: the man saw the girl.

Figure 3: Four Embedding Levels with Question Q0, targeting most deeply embedded structure

Level 2
Context: The teacher the student the driver saw hit is happy
Q: Who hit who?
A: the student hit the teacher.
Level 3
Context: The teacher the student the driver the girl saw hit likes is happy
Q: Who hit who?
A: the driver hit the student.
Level 4
Context: The teacher the student the driver the girl the man saw hit likes hates is happy
Q: Who hit who?
A: the girl hit the driver.

Figure 4: Embedding Levels 2-4 with Question Q1, targeting the next most deeply embedded structure

In figure 5 we present results for GPT-4 and GPT-3.5 for the four levels of embedding, with prompt P1 and question Q0. Both models are perfectly accurate for level 1 examples. Such examples tend to be very easy for humans. For GPT-3.5, accuracy falls sharply for levels 2 and 3, and is even lower for level 4. GPT-4 is far more accurate with higher levels of embedding – nearly perfect for levels 2 and 3, and still highly accurate (0.85) for level 4. This is striking, as these levels of embedding are not interpretable by human language users. Furthermore, multiple embeddings are almost certainly vanishingly rare in the training data for these models. In an extensive corpus study, Karlsson (2007)[p. 378] found that “in ordinary language use, written C3s [level 3] and spoken C2s [level 2] are almost non-existent”.

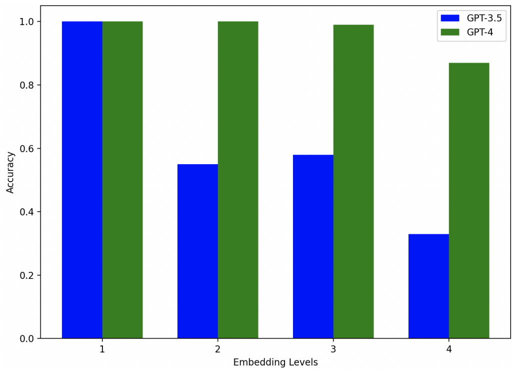

Figure 5: Accuracy of Center Embedding at levels 1-4, with Prompt P1 and Question Q0. GPT-4 is highly accurate even up to level 4, while GPT-3.5 is degraded at all levels above 1. (500 examples for each model, for each level)

4.1 Alternative Prompts and Questions

Model	P	Q	L1	L2	L3	L4
GPT-3.5	P0	Q0	0.86	0.49	0.34	0.14
GPT-3.5	P0	Q1	-	0.03	0.03	0.03
GPT-3.5	P1	Q0	1.00	0.55	0.58	0.33
GPT-3.5	P1	Q1	-	0.16	0.03	0.06
GPT-4	P0	Q0	1.00	0.55	0.28	0.11
GPT-4	P0	Q1	–	0.17	0.02	0.00
GPT-4	P1	Q0	1.00	1.00	0.99	0.87
GPT-4	P1	Q1	-	0.74	0.05	0.00

Table 1: Accuracy by Model and Embedding Level. (500 examples for each model, for each level)

In table 1 we present the accuracy of the two models with alternative prompt and question forms. In general, it is clear that both models are quite sensitive to these variations, in ways we are not in a position to explain. We would, however, like to draw attention to one specific observation: while the GPT-4 model achieves extremely high levels of accuracy with prompt P1 and question Q0, these levels drop precipitously with prompt P0 and question Q0, for all except level 1. We find this rather astonishing, since the only difference is that the model in the former case is provided with a single level 1 example, which is absent in the latter case. Somehow a single level 1 example has enabled GPT-4 to master higher levels of center embedding.

5 Conclusions

While multiple embedding structures are ubiquitous in human language, multiple center embeddings are different: they almost never occur, and are almost always uninterpretable for human language users. It has nonetheless been steadfastly maintained that they are grammatical, according to mainstream theories of human linguistic competence. In this paper, we have shown that GPT-3.5 struggles with center embeddings of any level greater than 1, much like humans, while GPT-4 performs very well with all four levels of center embeddings, thus apparently far exceeding human abilities. Why should this be?

One straightforward response is that GPT-4 is simply too big – at least with respect to its linguistic competence, the size of training data and number of system parameters is simply larger than needed, since it can now process linguistic structures that are far too complicated for humans.

There is another way to look at this, however. Chomsky famously argued that center embeddings are completely grammatical according to the theory of human linguistic competence. Humans, on this view, have a grammar that allows deeply embedded center embeddings, but this fact is obscured by performance factors – limitations on the general computational system in which the human language faculty is implemented. If the same linguistic competence could be implemented in a more powerful system, it would be easier to observe its true nature, since some of the performance limitations would be removed. Perhaps GPT-4 is just such a system: it has largely duplicated human linguistic competence, but is not subject to the same performance limitations as humans.

6 Limitations

The paper seeks to determine whether LLMs understand syntactic center embedding, but this general question is explored in only a few particular ways. First, only two LLMs are considered, and we suspect that other models might give quite different results. However, GPT-4 is the most powerful model we had access to, and we suspect that other less powerful models would, like GPT.3-5, have great difficulty with the tests reported on here. There are also several important limitations with respect to the data. First, the data is solely English. Second, it is synthetic data, constructed according to a template that reflects one specific form of center embedding, in which a noun phrase is modified by a relative clause. We believe this is the form of center embedding that is most familiar from the linguistics literature. However, there are other forms of center embedding that could also be considered. Furthermore, while we explored various combinations of different prompt and question forms, there are other forms and combinations that would be well worth exploring. Finally, we have made claims about the general uninterpretability of multiple center embeddings for humans; while these generally echo claims made in the literature, they are claims that would benefit from rigorous empirical examination.

References

GPT Models Overview. Accessed on 2023-12-10.

Noam Chomsky. 1957. *Syntactic structures*. The Hague: Mouton.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. 11. MIT press.

Noam Chomsky, George Armitage Miller, R Luce, R Bush, and E Galanter. 1963. Introduction to the formal analysis of natural languages. *1963*, pages 269–321.

Ruixiang Cui, Seolhwa Lee, Daniel Hershcovich, and Anders Søgaard. 2023. What does the failure to reason with “respectively” in zero/few-shot settings tell us about language models? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8786–8800, Toronto, Canada. Association for Computational Linguistics.

Daniel Hardt. 2023. Ellipsis-dependent reasoning: a new challenge for large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 39–47. Association for Computational Linguistics.

Fred Karlsson. 2007. Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, 43(2):365–392.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

A Appendix

A.1 Error Analysis

In all cases, the system is expected to produce answers of the form N1 V N2. We define three types of errors:

- Type 1: N1 is incorrect, N2 is correct
- Type 2: N1 is correct, N2 is incorrect
- Type 3: N1 is incorrect, N2 is incorrect

We consider selected settings based on a manual evaluation of the first 10 examples. Table 2 shows the percentage of errors of each type.

Model	**P**	**Q**	**L**	**T1**	**T2**	**T3**
GPT-3.5	P0	Q0	L1	0.00	0.00	1.00
GPT-3.5	P0	Q0	L2	0.10	0.90	0.00
GPT-3.5	P1	Q0	L2	0.00	0.90	0.10
GPT-3.5	P1	Q0	L3	0.00	0.90	0.10
GPT-3.5	P1	Q0	L4	0.00	0.90	0.10
GPT-4	P0	Q0	L2	0.00	0.90	0.10
GPT-4	P0	Q0	L3	0.00	0.90	0.10
GPT-4	P0	Q0	L4	0.00	0.80	0.20
GPT-4	P1	Q0	L4	0.00	0.90	0.10
GPT-4	P1	Q1	L2	0.40	0.00	0.60

Table 2: Error Types, T1, T2, T3 for selected settings of model, prompt type, question type and level of embedding (based on manual analysis of first 10 errors for each setting)

For all but two of the settings in table 2, nearly all the errors are of type T2, as in the following example:

> Context: The man the girl the driver knows hates is glad.
> Question: Who knows who?
> Model Answer: The driver knows the man.
> Correct Answer: The driver knows the girl.

Since the verb “knows” is explicit in the question, the model could simply assume that N1 is the noun phrase preceding “knows” in the context. This assumption ensures that a model avoids T1 errors, for question Q0. A T2 error arises in the above example, because the model selects “the man” rather than “the girl” as the second NP. Interestingly, GPT-3.5 has *only* T3 errors in the setting, P0, Q0, L1. In each case, it simply reverses N1 and N2, as in the following example:

Context: The woman the man hates left.
Question: Who knows who?
Model Answer: The woman hates the man.
Correct Answer: the man hates the woman.

Finally, GPT-4 has *only* T1 or T3 error types on the setting P1, Q1, L2. The following example illustrates a T3 error for this setting:

Context: The student the man the driver hates saw is glad.
Question: Who saw who?
Model Answer: The student saw the man.
Correct Answer: the man saw the student.

We have, of course, no direct insight into the strategies employed by these large language models in any of these settings. It seems intuitively plausible that models employ a strategy would normally get N1 right and N2 wrong, and this is indeed the pattern that arises with this limited error analysis. At this point we will offer no speculation about the two settings for which we observe different error patterns.

A.2 Sample Instantiations

We have the following substitutions for N and TV. N: (teacher, student, driver, girl, man), and TV: (saw, hit, likes, hates, knows). IV is always substituted with the phrase, "is happy".