Linguistic Competence and Performance for LLMs: the Case of Syntactic Center Embedding

Anonymous ACL submission

Abstract

We consider 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, this is generally not the case for multiple center embeddings. Despite this, it is a standard view in linguistic theory that multiple center embeddings are grammatically acceptable – human linguistic competence includes this ability, but this is obscured by performance limitations. We construct sentences with center embeddings of varying levels, ranging from 1-4, and we find that GPT-4 achieves nearly perfect results even with 3 or 4 levels of embeddings. Other LLMs show a sharp drop in accuracy above level 1. We suggest that this is because GPT-4 has successfully learned the same underlying linguistic competence as humans, while not being subject to the same performance limitations. This would mean that human linguistic competence is more clearly observed in GPT-4 than in humans.

1 Introduction

Until recently, there was a simple reason why every AI system would fail the Turing Test – they lacked the basic linguistic capabilities shared by all native speakers of a language. That has changed with current large language models (LLMs), which, it would seem, have now mastered human language. As Mahowald et al. (2024)[p. 2] put it, "for LLMs starting with GPT-3, their formal [linguistic] competence is essentially at ceiling". There remain, however, notable differences in the linguistic behavior of LLMs and humans. In this paper we focus on differences in the interpretation of syntactic center embedding constructions. These constructions, while little noted in the NLP literature, have a special significance in the development of modern linguistics. Famously, Chomsky claims that center embedding is fully grammatical as a matter of linguistic competence, but generally fails to be accepted because of a performance limitation involving short-term memory (Chomsky, 1957; Chomsky et al., 1963). These claims are central to the very founding of modern linguistics.

It is revealing to compare center embedded constructions to left and right embedding. Consider a propositional verb like "believe", that can take a sentence as its complement to the 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, both to the right and to the left. We turn now to center embedding. Here the embedding clause contains material both to the left and right of the embedded clause. This is illustrated by (3), where 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

[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.

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.

In this paper, we explore whether LLMs can interpret center embedding structures. We find that GPT-4 performs extremely well at all levels, from 1 to 4, while other, less powerful models, exhibit behavior that is rather similar to humans, performing well at level 1, but quite poorly at higher levels. This is an apparent paradox, since it is the less powerful models that more closely correspond to human linguistic behavior. We suggest that the Chomskyan distinction between competence and performance provides a resolution of this paradox – the more powerful GPT-4 model has successfully learned human linguistic competence, but is not subject to the same performance limitations as humans. Indeed, it may be that human linguistic competence can be observed more clearly in GPT-4 than in humans.

2 Related Work

2.1 Center Embedding and Linguistic Competence

According to Karlsson (2007, p. 365) "the mainstream view...voiced by many linguists from different camps" is that "there are no grammatical restrictions on multiple center-embedding of clauses." 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, 1965)[p. 4]. Performance factors, such as memory limitations, 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.

2.2 Center Embedding and Performance Factors

Gibson (1998)[p. 3] notes that center embedding structures give rise to what is often "referred to as a *processing overload* effect". Gibson proposes the Syntactic Prediction Locality Theory (SPLT). According to this theory, center embedding incurs a memory cost, associated with "computational resources [that] are required to store a partial input sentence" (Gibson (1998)[p. 8]). This is an essential feature of center embedding constructions; for example, in (4) above, when the word "driver" is encountered, there are three partial input sentences that must be stored. On this theory, it is the requirement to keep multiple partial structures in memory that can lead to processing overload. Gibson (1998)[p. 14] observes that this "... fits with what is known about short-term memory recall in non-linguistic domains: it is harder to retain items in short-term memory as more interfering items are processed."

Gibson considers a wide range of differences in types of embedding structures in arguing for the superiority of SPLT over previous theories, such as Chomsky et al. (1963), Miller and Isard (1964), Abney and Johnson (1991), and Engelmann and Vasishth (2009). What Gibson's theory shares with the previous theories is the view that the facts about center embedding structures are explained with reference to performance factors.

2.3 Corpus Study

Karlsson (2007) reports on a study of "corpus data from seven Standard Average European (SAE) languages: English, Finnish, French, German, Latin, Swedish, and Danish." Karlsson shows that level 2 center embeddings, while rare, do occur; overall Karlsson (2007)[p. 378] finds that "in ordinary language use, written C3s [level 3] and spoken C2s [level 2] are almost non-existent".

2.4 Linguistic Probing of LLMs

There is an extensive literature describing the probing of LLMs for specific linguistic capabilities. Mahowald et al. (2024) argue that current LLMs have largely mastered what they call "formal linguistic competence". They point out that current models perform well on resources such as the BLiMP benchmark (Warstadt et al., 2020), which consists of minimal pairs illustrating many linguistic phenomena. "Models achieve similarly impressive results," they continue, "on other linguistic benchmarks like SyntaxGym" (Gauthier et al., 2020).

However, some recent works have shown that there remain specific capabilities that pose difficulties for some of the most powerful current models. For example Hardt (2023) shows that recent LLMs struggle with the phenomenon of ellipsis while Cui et al. (2023) find that they have substantial difficulties interpreting sentences with "respectively".

2.5 Human Performance

We posed 4 examples each of levels 1, 2 and 3, to 9 respondents, for a total of 108 observations. The context and question were modeled after the materials used in our LLM experiments.[2] As shown in table 2 the results show a sharp drop in accuracy from level 1 to levels 2 and 3; consistent with widely held views in the literature.

Level	Accuracy
1	.889
2	.611
3	.528

Table 1: Survey Results for Center Embeddings

There are numerous empirical studies that further support the claim that center embedding presents difficulties for humans, although there is considerable methodological variety. For example, Thomas (1995)[p. 22] asks subjects to rate examples according to perceived difficulty "on a quick first reading". Thomas shows that there are important differences based on the type of center embedding. However, in general, he notes that a simple level 1 structure "is easy to understand", while "embedding just one more clause [i.e. level 2]... produces near incomprehensibility." (Thomas (1995)[p. 8]) Bach et al. (1986) describe a psycholinguistic study concerning somewhat different embedding constructions in German and Dutch, again finding a striking difference in difficulty between level 1 and higher levels of embedding.

[2]Survey data provided online upon acceptance.

3 Data

We construct a synthetic dataset, where each item consists of a context, a prompt and a question.[3] We consider each of these elements in turn.

3.1 Context

The context consists of synthetic examples of center embedding of levels 1-4. 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 appendix A.3 for instantiations of N, TV, and IV.)

3.2 Prompt

We define the prompt shown in figure 1, which we designate as P1. The prompt includes a single example, exhibiting level 1 center embedding.

```
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 – single example

3.3 Question

We formulate a question, "Who TV'ed who", where the verb TV is from the most deeply embedded clause. This question is Q0 (figure 2).

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

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 2: Four Embedding Levels with Question Q0, targeting the most deeply embedded structure

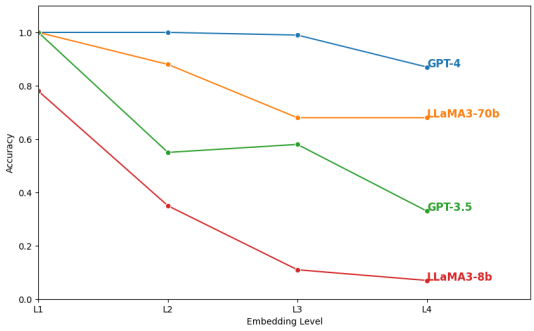

Figure 3: Accuracy of Center Embedding at levels 1-4, with Prompt P1 and Question Q0.

4 Test

4.1 Initial Test

For each embedding level (1-4), we construct 500 synthetic examples, and we test four models: GPT-3.5, GPT-4, Llama3-70B and Llama3-8B (see Appendix A.2 for details). Our initial test uses prompt P1 and question Q0. In figure 3 we present results. For level 1 examples, all models are perfectly accurate except for the smaller Llama3-8b. GPT-4 remains nearly perfect at higher levels, while GPT-3.5 has a sharply lower accuracy at levels higher than 1. In this way, it mirrors human behavior, while Llama3-70b occupies an intermediate position, with a less precipitous drop between levels 1 and 2.

According to the competence model discussed above, center embeddings are fully grammatical at any level. GPT-4 seems closely aligned with the competence model. However, the fact that accuracy drops at level 4 suggests that GPT-4 does not perfectly reflect the competence model. We return to this point below in examining alternative prompts and questions.

4.2 Alternative Prompts and Questions

We define two alternative prompts: P0, which lacks an example (figure 4), and P2, which contains two examples – a level 1 example and a level 2 example (figure 5). We also define a question variant, Q1, which targets the next most deeply embedded predication, as exemplified in figure 6.

We have, then, prompts P0, P1, and P2, and questions Q1, and Q2. This gives 6 settings for each of our 4 models, for a total of 24 results. The complete results are given in table 3, in appendix A.4. We draw attention to the results with GPT-4. Above, we argued that GPT-4 quite closely matched the competence model, based on the results shown in figure 3. This is based on Q0, which targets only the most deeply embedded predication. With Q1, GPT-4 shows a much lower accuracy on levels 2-4, with prompt P1. This conflicts with the suggestion that GPT-4 is reflecting the competence model; something we noted above with respect to the drop in accuracy on level 4 with Q0. On the other hand, with prompt P2, the accuracy of GPT-4 increases on question Q1, although still not to the nearly perfect level of accuracy seen with Q0.

5 Conclusions

In this paper, we found that GPT-4 interprets multiple center embeddings with high accuracy, while less powerful models struggle with them, just as humans do. We have suggested that GPT-4 has successfully learned human competence grammar, while not being subject to the same performance limitations as humans. On this view, human linguistic competence would be more clearly observable in a powerful LLM than it is in humans. This may or may not be the right explanation. But it is offered as a first step in elaborating a theory that takes account of both the competence and the performance of LLMs.

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 four LLMs are considered, and we suspect that other models might give quite different results. 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

Steven P Abney and Mark Johnson. 1991. Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, 20:233–250.

Emmon Bach, Colin Brown, and William Marslen-Wilson. 1986. Crossed and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes*, 1(4):249–262.

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

Noam Chomsky. 1965. *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.

Felix Engelmann and Shravan Vasishth. 2009. Processing grammatical and ungrammatical center embeddings in English and German: A computational model. In *Proceedings of the Ninth International Conference on Cognitive Modeling, Manchester, UK*, pages 240–45.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

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. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

George A Miller and Stephen Isard. 1964. Free recall of self-embedded english sentences. *Information and control*, 7(3):292–303.

James David Thomas. 1995. *Center-embedding and self-embedding in human language processing*. Ph.D. thesis, Massachusetts Institute of Technology.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

A Appendix

A.1 Error Analysis

In all cases, the system is expected to produce answers of the form N1 V N2. We define four 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
- Type 4: Verb is incorrect, or answer is not of the required form

We consider selected settings based on a manual evaluation of the first 10 examples. We focus on Level 2, with Prompt P1 and Q0 for the models GPT.3-5, Llama3-8b and Llama3-70b. All three of these models have a substantial number of errors here. On the other hand, GPT-4 has no errors in this case. Instead we look at Q1 with GPT-4. Table 2 shows the number of errors of each type.

Model	P	Q	T1	T2	T3	T4
GPT-3.5	P1	Q0	0	9	1	0
Llama3-8b	P1	Q0	0	7	2	1
Llama3-70b	P1	Q0	0	7	3	0
GPT-4	P1	Q1	4	0	6	0

Table 2: Error Types, T1, T2, T3, and T4 for selected settings of model, prompt type, and question type (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 LLMs in any of these settings. It seems intuitively plausible that models employ a strategy that 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 Models

The GPT-4 and GPT-3.5-turbo models were accessed from the OpenAI site in the period from 10 June 2024 to 21 June 2024, with default settings. The Llama3-70b and Llama3-8b models were accessed from api.llama-api.com in the same period, also with default settings.

A.3 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".

A.4 Complete Results

Results for all models and settings are given in table 3.

A.5 Alternative Prompts and Questions

> 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 4: Prompt P0 – no samples

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-3.5	P2	Q0	1.00	0.59	0.69	0.32
GPT-3.5	P2	Q1	–	0.38	0.08	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
GPT-4	P2	Q0	1.00	1.00	1.00	0.97
GPT-4	P2	Q1	–	0.94	0.51	0.13
Llama3-8b	P0	Q0	0.96	0.46	0.17	0.08
Llama3-8b	P0	Q1	–	0.04	0.06	0.03
Llama3-8b	P1	Q0	0.78	0.35	0.11	0.07
Llama3-8b	P1	Q1	–	0.07	0.02	0.02
Llama3-8b	P2	Q0	0.86	0.49	0.20	0.12
Llama3-8b	P2	Q1	–	0.10	0.04	0.03
Llama3-70b	P0	Q0	0.98	0.65	0.67	0.48
Llama3-70b	P0	Q1	–	0.05	0.06	0.03
Llama3-70b	P1	Q0	1.00	0.88	0.68	0.68
Llama3-70b	P1	Q1	–	0.08	0.04	0.04
Llama3-70b	P2	Q0	1.00	0.89	0.67	0.71
Llama3-70b	P2	Q1	–	0.20	0.11	0.02

Table 3: Accuracy by Model and Embedding Level. (500 examples for each model, for each level)

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 are two samples:

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

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

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

Figure 5: Prompt P2 – two samples

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