Sparks of Pure Competence in LLMs: the Case of Syntactic Center Embedding in English

Anonymous ACL submission

Abstract

Linguistic theory distinguishes between competence and performance: the competence grammar ascribed to humans is not always clearly observable, because of performance limitations. This raises the possibility that an LLM, if it is not subject to the same performance limitations as humans, might exhibit behavior closer to a pure instantiation of the human competence model. We explore this in the case of syntactic center embedding, where, the competence grammar allows unbounded center embedding, although humans have great difficulty with any level above one. We study this in four LLMs, and we find that the most powerful model, GPT-4, does appear to be approaching pure competence, achieving high accuracy even with 3 or 4 levels of embeddings, in sharp contrast to humans and other LLMs.

> "The heptapods had no objection to the center-embedding of clauses, something that quickly defeated humans".
>
> – Story of Your Life (Chiang, 1998)

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 embedding with 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]

[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.

(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 are generally uninterpretable for human language users, and are virtually nonexistent in normal texts.

In this paper, we explore whether LLMs can interpret and assess center embedding structures in English. We create synthetic data instantiating levels 1-4, and pose questions which require understanding of the structure. For example, for example (4) above, we ask, "Who hit who?", a question that targets the most deeply embedded predication. Here, we find that GPT-4 performs extremely well at all levels, from 1 to 4, in contrast to other models, and also sharply contrasting with what is known about human behavior. This, we argue, suggests that GPT-4 is approaching pure competence. We perform a total of four different tests, varying the embedding level that is questioned, the number of few-shot learning examples provided, and the lengths of NPs in the synthetic data. We also test the ability to assess the grammaticality of center embedding structures.

The results of these additional tests are mixed. On the one hand, in all the tests, there are settings in which GPT-4 performs with very high accuracy, suggesting something close to a pure instantiation of the competence model. On the other hand, there are also tasks and settings in which its performance is degraded, revealing sensitivity to factors such as the embedding level of the question, the number of few-shot examples, and the lengths of the NPs in the structures.

In light of these mixed results, it is premature to conclude that we can observe pure competence in an LLM like GPT-4. Yet its behavior is much closer to pure competence than human behavior. We discuss the implications of this, noting that GPT-4 has attained these impressive abilities, despite the fact that multiple center embeddings are almost entirely non-existent in its training data. We conclude with some reflections about the implications of these results for theorizing about the language faculty as it is instantiated in humans as well as in AI models.

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." This is all the more striking given the extreme rarity of multiple center embedding. Karlsson (2007, p. 378) reports on a study of "corpus data from seven Standard Average European (SAE) languages: English, Finnish, French, German, Latin, Swedish, and Danish", finding that "in ordinary language use, written C3s [level 3] and spoken C2s [level 2] are almost non-existent."

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). Millière (2024) points out that "Linguistic performance can be affected by external factors like memory limitations, distractions, slips of the tongue, etc. that may obscure the full extent of the underlying competence." Performance factors make the underlying linguistic competence difficult to observe in humans, 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), and Abney and Johnson (1991). 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 Human Performance

There are numerous empirical studies that support the claim that center embedding presents difficulties for humans. 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. We performed a small, informal survey to further examine human performance on center embedding. See A.2 for details.

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".

2.4.1 Subject-Verb Agreement

A particular area of interest for linguistic probing is subject-verb agreement. Wilson et al. (2023, p. 278) point out that subject-verb agreement "depends not on linear proximity to the verb, but structural proximity ...", as illustrated by the following paradigm:

(8) a. The labels on the bottle is ...
b. * The label on the bottle is ...
c. * The labels on the bottle are ...
d. The label on the bottle are ...

Humans sometimes diverge from the pure competence model, making errors based on an "attractor", i.e., a noun that intervenes between subject and verb, such as "bottle" in example (8)b above. Recent work (Linzen et al., 2016; Lakretz et al., 2021) has shown that models are able to largely capture the "structure-sensitive grammatical knowledge" implicated in the competence model (Wilson et al., 2023, p. 278), while also showing some errors based on attractor effects.

2.4.2 Center Embedding

Just as with subject-verb agreement, human performance diverges from the competence model with center embedding. However, the divergence is much starker in the case of center embedding – humans consistently fail in the interpretation of multiple center embeddings, although they are completely acceptable according to the competence model. Recent probing of LLMs reveals similar divergence from the competence model. For example Dentella et al. (2023) find that LLM "accuracy on grammatical prompts of center-embedded sentences is at chance" in a test of grammatically judgments by LLMs in the GPT-3 family. Hu et al. (2024, p. 10) test LLMs on a variety of constructions, finding that models "evaluated on the same sentences in minimal pairs achieve at- or near-ceiling performance on most linguistic phenomena tested, except for centre embedding", noting that,

for center embedding, "humans also perform near chance."

An additional observation comes from Gibson and Thomas (1999), concerning what they call the "VP illusion", where ungrammatical versions of center embedding sentences are judged to be as acceptable as their grammatical counterparts, as illustrated by (9):

(9) a. The patient who the nurse who the clinic had hired met Jack.
b. The patient who the nurse who the clinic had hired admitted met Jack.

Example (9)b is a grammatical level 2 example of center embedding, while (9)a is ungrammatical, since the verb "admitted" is omitted. Gibson and Thomas find that the ungrammatical examples with a missing VP, like (9)b, are judged to be as acceptable as their grammatical counterparts. Subjects were given seven "practice examples", with "discussion of possible scores for each" (Gibson and Thomas, 1999, p. 238). The study was performed using a questionnaire, and the authors note that, although subjects were instructed to read examples only a single time, subjects had the opportunity to re-read examples. Christiansen (1997) reports on a variant of this study where examples are presented online, so that re-reading is not possible. In this experiment, the missing VP examples were perceived as more acceptable than their grammatical counterparts. See also Engelmann and Vasishth (2009) for an alternative view, arguing that the illusion does not arise for German speakers.

3 Data

We construct a synthetic dataset, where each item consists of a prompt, a context, and a question.[2] 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 denotes Noun, TV denotes Transitive Verb and IV denotes 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.

We have the following substitutions for N and TV:

- **N**: teacher, student, driver, girl, man, woman, boy
- **TV**: saw, hit, likes, hates, knows
- **IV**: is happy, left, is glad

The synthetic data is constructed for levels 1-4, by a procedure that repeatedly makes random selections for N, TV, and IV, resulting in a large collection of sentences for each level. For each test, a random subset of unique sentences are selected.

3.2 Prompt

We define the prompt P0, shown in figure 1. We also use prompts with examples, thus applying few-shot learning. The examples within the prompt are always of the same embedding level as the example in the context.

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 noun, and V stands for a verb. Context: {context} Question: {question} Now answer the question:

Figure 1: Prompt P0

We will use prompts with varying numbers of few-shot examples, such as P5, P10 and P20, i.e., with 5, 10 and 20 few-shot examples respectively.

3.3 Question

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

We also define a question, Q1, that targets the next most deeply embedded predication, as exemplified in figure 3. Note that Q1 does not apply to level 1 examples.

We evaluate the model response as correct if it matches the predefined answer exactly, and incorrect otherwise. All tests use accuracy as the metric.

[2] 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

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

4 Testing

4.1 Test 1: Question Q0

For each embedding level (1-4), we test four models: GPT-3.5, GPT-4, Llama3-70B and Llama3-8B (see Appendix A.1 for details). Test 1 uses question Q0, with either 0 or 5 few-shot examples. In table 1 we present results. GPT-4 is perfect at level 1 with both few-shot settings. With 0 examples, accuracy declines rapidly with higher embedding levels, while with 5 examples, GPT-4 continues to have very high accuracy up to level 4. The other models all have much lower accuracy than GPT-4, especially with higher embedding levels. According to the competence model, center embeddings are fully grammatical at any level. With 5 few-shot examples, GPT-4 seems closely aligned with the competence model, although there is a modest drop in accuracy at levels 3 and 4. The other three models are more similar to humans, in that they have considerable difficulty with any multiple levels of embedding.

4.2 Test 2: Question Q1

In Test 2, we pose question Q1, and we use prompts with few-shot examples, ranging from 0 to 30. One interpretation of the test 1 results is that GPT-4 with 5 examples is indeed approaching pure competence with respect to center embedding, with nearly perfect results up to level 3, and still quite high results with level 4, contrasting sharply with humans and the other LLMs. On the other hand, it could be that the behavior of GPT-4 does not actually reflect the competence model involving unbounded structural embedding; there are other conceivable explanations. It could, for example, be employing a simple linear strategy, where it conducts a search to the left of the verb being questioned to locate the subject and object NP's. Consider the example in figure **??**. When posed with the question "Who saw who?", the strategy might be to locate the two NP's immediately to the left of "saw". The first NP encountered is the subject, and the second is the object. This strategy is perhaps facilitated by the fact that all NPs in our synthetic data consist of two words.

By using question Q1, we seek to rule out a linear strategy along the lines given above. Consider the level 2 example in figure 3. To answer the question, "Who hit who?", it is necessary to search left by first skipping over the verb "saw" and the NP "the driver". While this is not inconceivable, it would seem to be more complicated than is the case with question Q0. In test 2 we also experiment with the number of examples in few shot learning, using prompts with up to 30 few-shot examples.

The results are given in table 2. The llama models struggle with Q1, even at level 2. Here GPT-3.5 also struggles, although accuracy does increase markedly as the number of few-shot examples increases. Things are quite different with GPT-4 –

Model	Few-shot	L1	L2	L3	L4
llama3-8b	0	0.005	0.005	0.000	0.000
llama3-8b	5	0.005	0.005	0.005	0.015
llama3-70b	0	0.845	0.640	0.535	0.455
llama3-70b	5	0.760	0.465	0.210	0.095
gpt-3.5-turbo	0	0.545	0.355	0.110	0.045
gpt-3.5-turbo	5	**1.000**	0.885	0.580	0.315
gpt-4	0	**1.000**	0.500	0.385	0.195
gpt-4	5	**1.000**	**1.000**	**0.900**	**0.845**

Table 1: Test 1 – Question Q0, Accuracy levels 1-4

while it encounters some difficulty with Q1 as compared with Q0, accuracy increases sharply with few-shot examples. Already with 5 examples, GPT-4 is above .9 for levels 2 and 3, and with 25 examples it achieves a score of .840 on level 4.

Model	Few-shot	L2	L3	L4
llama3-8b	0	0.000	0.000	0.000
llama3-8b	5	0.000	0.000	0.000
llama3-8b	10	0.000	0.000	0.000
llama3-8b	20	0.000	0.000	0.000
llama3-70b	0	0.040	0.035	0.040
llama3-70b	5	0.200	0.225	0.010
llama3-70b	10	0.115	0.175	0.130
llama3-70b	20	0.175	0.145	0.000
GPT-3.5	0	0.000	0.000	0.005
GPT-3.5	5	0.565	0.205	0.160
GPT-3.5	10	0.710	0.365	0.075
GPT-3.5	20	0.645	0.325	0.245
GPT-3.5	25	0.870	0.565	0.350
GPT-3.5	30	0.795	0.525	0.315
GPT-4	0	0.165	0.015	0.000
GPT-4	5	0.905	0.980	0.410
GPT-4	10	0.950	0.980	0.335
GPT-4	20	**1.000**	**1.000**	0.435
GPT-4	25	0.995	**1.000**	**0.840**
GPT-4	30	0.995	**1.000**	0.690

Table 2: Test 2 – Question Q1

4.3 Test 3: Variable-Length NPs

In test 3, we create an additional difficulty for the kind of linear strategy discussed above. We modify the test data so that NP's are sometimes two words, and other times three words. This is done by modifying the instantiations for N as follows:

N: happy teacher, young student, driver, girl, man, woman, short boy

Recall that, in our synthetic data, all transitive verbs consist of a single word, and all NP's consist of two words. So, if we consider again the level 2 example in figure 3 with the Q1 question, "Who hit who?" a conceivable search strategy would be: search 4 words to the left, at which point the subject and object NP's are encountered. With variation in the lengths of NPs, a strategy of searching left can no longer be determined by the number of words encountered. Rather, such a strategy would have to be defined in terms of constituents. Results are shown in figure 3. Only GPT-3.5 and GPT-4 are tested here, since the llama models performed so poorly in test 2. It does appear that the variable length of NP's poses an additional challenge for the models. However, similarly to test 2, accuracy rises sharply as few-shot examples increase.

Model	Few-shot	L2	L3	L4
GPT-3.5	0	0.005	0.030	0.015
GPT-3.5	5	0.450	0.270	0.060
GPT-3.5	10	0.710	0.325	0.175
GPT-3.5	15	0.745	0.295	0.090
GPT-3.5	20	0.670	0.285	0.200
GPT-4	0	0.045	0.010	0.005
GPT-4	5	0.995	0.740	0.260
GPT-4	10	0.915	0.830	0.150
GPT-4	15	**0.950**	**1.000**	**0.635**
GPT-4	20	0.870	0.990	0.600

Table 3: Test 3 – Question Q1, variable-length NPs

4.4 Test 4: Missing VP Illusion

In test 4, the model is prompted to judge whether an example is grammatically correct or not. Here we

restrict attention to GPT-4. Half of the examples are taken from our original synthetic data, as described above for test 1. We create an equal-sized set of examples with a missing verb, as illustrated for level 2, by (10):

(10) a. *The teacher the student the driver saw is happy.
b. The teacher the student the driver saw hit is happy.

We test with data for levels 2, 3 and 4. The accuracy of judgments is at or below chance (.50) for few-shot values of 0 or 5. However, with few-shot of 10, GPT-4 is performing notably better than humans, well above chance for all three levels. Note that, in the study of Gibson and Thomas (1999), subjects were given 7 "practice examples". Furthermore, they were only tested on level 2 examples.

Model	Few-shot	L2	L3	L4
GPT-4	0	0.405	0.410	0.495
GPT-4	5	0.485	0.525	0.460
GPT-4	10	**0.835**	**0.665**	**0.590**

Table 4: Test 4 – Missing Verb Grammaticality Judgment

4.5 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: Other

We consider selected settings based on a manual evaluation of the first 10 examples, restricting attention to gpt-4, in test 1 and test 2. Table 6 shows the number of errors of each type. While there is considerable variation, some clear tendencies can be observed in this small-scale error analysis. With Q0, errors tend to be Type 2, which might relate to the fact that the subject, N1, is adjacent to the verb being questioned. This might explain the comparative lack of errors with N1 for Q0. This is not the case with Q1, and here both type 1 errors and type 3 errors are frequent.

Model	Level	Few-shot	Q	T1	T2	T3
GPT-4	2	0	Q0	0	10	0
GPT-4	3	0	Q0	0	9	1
GPT-4	4	0	Q0	0	9	1
GPT-4	2	5	Q0	10	0	0
GPT-4	3	5	Q0	0	10	0
GPT-4	4	5	Q0	0	10	0
GPT-4	2	0	Q1	0	1	9
GPT-4	3	0	Q1	2	0	8
GPT-4	4	0	Q1	8	0	2
GPT-4	2	5	Q1	10	0	0
GPT-4	3	5	Q1	2	0	2
GPT-4	4	5	Q1	0	7	3

Table 5: Error Types, T1, T2, T3, and T4 for selected test settings, based on manual analysis of first 10 errors for each setting

5 Discussion

Chomsky (1965, p. 4) describes competence as a theory of the "mental reality underlying actual behavior". As with any domain of natural phenomena, there are an unbounded number of potential theories that are consistent with observation, so other factors, such as elegance and simplicity, play a key role in selecting among candidate theories (Kuhn, 1997). According to the Chomskyan framework, the theory of linguistic competence is formulated in terms of simple recursive rules. While this model sometimes deviates from observed linguistic behavior, these deviations can plausibly be attributed to performance factors.

Dupre (2021, p. 632) notes that, on mainstream views in linguistics, "the gap [between competence and performance] is quite substantial", and thus finds it unlikely that an LLM would "provide insight . . . to linguistic competence." Yet this is the conclusion we argue for in this paper – that linguistic competence can be more clearly observed in GPT-4 than in humans.

The evidence for this conclusion has been presented in tests 1-4 described above, and can be largely summarized in figure 4. Here we can see that there are certain settings in which GPT-4 maintains high accuracy in multiple embeddings. In this way GPT-4 differs sharply with the other, less powerful models we tested, and of course this is also quite different from what is observed with human performance.

The evidence we have presented is far from con-

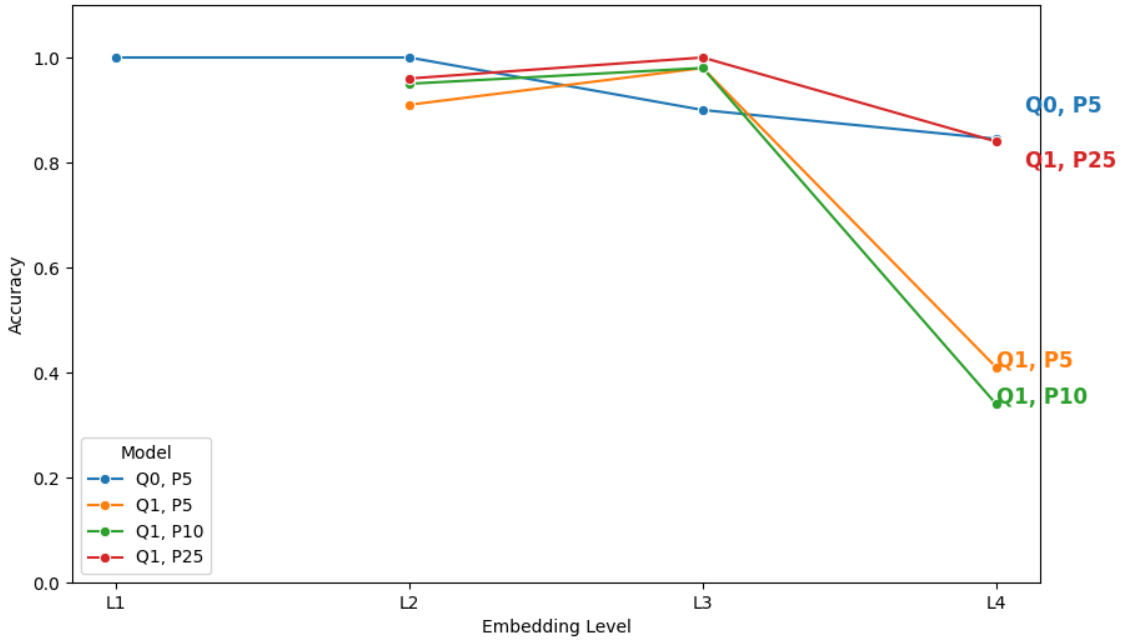

Figure 4: GPT-4 Results, tests 1 and 2
(L1 is not relevant for question Q1)

clusive. Even in the best settings, such as Q0, P5 and Q1, P25, the accuracy is not perfect, and furthermore declines notably at level 4. Our tentative explanation is that, while GPT-4 may well have acquired the linguistic competence model, it is also subject to certain performance limitations, although these limitations are far less severe than those that apply to humans. Another important issue involves few-shot learning. GPT-4 does not achieve high accuracy in the zero-shot setting. It could be argued that GPT-4 does not in fact implement the competence model, but rather, is simply exhibiting effective few-shot learning. We have a different view, based on the idea that it can be difficult to access the knowledge of an LLM through prompt-based tasks. As Hu and Levy (2023, p. 9) argue, "A model's failure to exhibit a linguistic generalization when prompted might not reflect a lack of the relevant information"; Hu and Frank (2024, p. 1) note, furthermore, that "performance on a task is a function of the model's underlying knowledge, combined with the model's ability to interpret and perform the task." In our view, the provision of the few-shot learning examples serves to support the model's "ability to interpret and perform the task", thus providing a more accurate reflection of the underlying competence of the model.

6 Conclusions

In this paper, we have explored the possibility that a powerful LLM might reflect pure competence. That is, it might faithfully reflect the human competence model. In humans, linguistic competence is often obscured by performance limitations. Center embeddings present perhaps the most striking divergence between human linguistic behavior and the competence model. We report on a series of tests involving a variety of settings of few-shot learning, embedding levels, and constituent sizes, as well as a grammaticality judgment test. The results are mixed, in that GPT-4 performs very well in many, but not all, settings. We suggest that GPT-4 might be subject to less strict performance limitations than humans, so that competence is less obscured by performance limitations in GPT-4 than it is in humans.

7 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. There are also several important limitations with respect to the data. First, the data is solely English. Sec-

ond, 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. There are other forms of center embedding that could also be considered. Furthermore, while we explored various combinations of different question types, few-shot learning, and constituent lengths, 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.

Ted Chiang. 1998. Story of your life. *Stories of your life and others*, pages 117–78.

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.

Morten H Christiansen. 1997. The (un) grammaticality of doubly center-embedded sentences: a connectionist perspective. In *Poster presented at the 10th CUNY Sentence Processing Conference, Santa Monica, CA*.

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.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Gabe Dupre. 2021. (what) can deep learning contribute to theoretical linguistics? *Minds and Machines*, 31(4):617–635.

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.

Edward Gibson and James Thomas. 1999. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14(3):225–248.

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.

Jennifer Hu and Michael C Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. *arXiv preprint arXiv:2404.02418*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. *arXiv preprint arXiv:2305.13264*.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

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

Thomas S Kuhn. 1997. *The structure of scientific revolutions*, volume 962. University of Chicago press Chicago.

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699. Special Issue in Honour of Jacques Mehler, Cognition's founding editor.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

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.

Raphaël Millière. 2024. Language models as models of language. *arXiv preprint arXiv:2408.07144*.

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.

Michael A Wilson, Zhenghao Zhou, and Robert Frank. 2023. Subject-verb agreement with seq2seq transformers: Bigger is better, but still not best. *Proceedings of the Society for Computation in Linguistics*, 6(1):278–288.

A Appendix

A.1 Test Details

A.1.1 Test 1

The GPT-4 and GPT-3.5-turbo models were accessed from the OpenAI site in the period from 13 to 14 December 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. Each of these tests were performed with 200 randomly selected examples.

A.1.2 Test 2

The GPT-4 and GPT-3.5-turbo models were accessed from the OpenAI site in the period from 10 November 2024 to 1 December 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. Each of these tests were performed with 200 randomly selected examples.

A.1.3 Test 3

The GPT-4 and GPT-3.5-turbo models were accessed from the OpenAI site in the period from 10 November 2024 to 1 December 2024, with default settings. Each of these tests were performed with 200 randomly selected examples.

A.1.4 Test 4

The GPT-4 model was accessed from the OpenAI site in the period from 10 November 2024 to 1 December 2024, with default settings. Each of these tests were performed with 200 randomly selected examples.

A.2 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.[3] As shown in table 6 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 6: Survey Results for Center Embeddings

[3] Survey data provided online upon acceptance.