



Can GPT-4o Reason about Language? A Syntax Challenge

Hye-Won Choi (Ewha Womans University) · Soo-Yeon Kim (Sejong University) · Sanghoun Song (Korea University)



This is an open-access article distributed under the terms of the Creative Commons License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received: September 11, 2025

Revised: October 15, 2025

Accepted: October 25, 2025

Choi, Hye-Won (First author)
Professor, Department of English Language and Literature, Ewha Womans University
52, Ewhayeodae-gil, Seodaemun-gu, Seoul, 03760, Republic of Korea
Email: hwchoi@ewha.ac.kr

Kim, Soo-Yeon (Co-author)
Professor, English Data Convergence Major, Sejong University
209, Neungdong-ro, Gwangjin-gu, Seoul, 05006, Republic of Korea
Email: kimsy@sejong.ac.kr

Song, Sanghoun (Corresponding author)
Associate Professor, Department of Linguistics, Korea University
145 Anam-ro, Seongbuk-gu, Seoul, 02841, Republic of Korea
Email: sanghoun@korea.ac.kr

ABSTRACT

Choi, Hye-Won, Soo-Yeon Kim and Sanghoun Song, 2025. Can GPT-4o reason about language? A syntax challenge. *Korean Journal of English Language and Linguistics* 25, 1468-1495.

This study investigates the capacity of GPT-4o, a multimodal large language model, to engage in linguistic analysis through a syntax exam designed to probe foundational concepts such as constituency, ambiguity, and recursion in both English and Korean. While often providing accurate definitions and fluent responses, the model struggles to apply syntactic principles consistently, especially in tasks requiring structural reasoning and tree diagram generation. The model's frequent misinterpretations and incoherent analyses within and across tasks reveal a reliance on pattern recognition and heuristics rather than a systematic grasp of hierarchical structures and fundamental linguistic reasoning. These findings point to the limitations of current large language models in performing metalinguistic analyses, exposing a gap between surface-level performance and genuine metalinguistic competence, which in turn presupposes linguistic competence. By examining GPT-4o's responses across a range of syntactic challenges, this study emphasizes the need for more rigorous evaluation frameworks that go beyond surface-level fluency to assess models' capacity for human-like linguistic reasoning and analysis.

KEYWORDS

Large Language Model (LLM), GPT-4o, syntax, constituency, ambiguity, recursion

1. Introduction

Recent large language models (LLMs), including GPT models, demonstrate impressive linguistic performance, appearing to possess a surprisingly robust command of human language and the ability to learn grammatical features such as anaphora, filler-gap dependencies, island constraints, long-distance dependency, NPIs, scope ambiguities, etc. (Chang and Bergen 2024, Howitt et al. 2024, Kamath et al. 2024, Kobzeva et al. 2025, Wilcox et al. 2024, among others). However, since behavioral linguistics tests—such as evaluating grammaticality or performing syntactic operations—assess only language performance (Haider 2023), it remains questionable whether such performance reflects genuine linguistic ability or merely reproduces surface patterns without deeper understanding (Bender et al. 2021, Mahowald et al. 2024, among others). Shin et al. (2023), for instance, examined language models' sensitivity to negative polarity item licensing and found partial alignment with human judgments, likely driven by statistical co-occurrence rather than syntactic reasoning.

Building on this line of research, the present study investigates GPT-4o's metalinguistic reasoning across a range of syntactic phenomena, thereby offering a complementary view of the limits of structural understanding in current LLMs. GPT-4o, a multimodal large language model developed by OpenAI and released in May 2024, features capabilities in multilingual processing and multimodal input including text, image, and audio.¹ This study directly examines GPT-4o model's grasp of core linguistic knowledge through a syntax exam, on the premise that analyzing language in structural terms reveals how the model draws on linguistic principles to guide its surface behavior. Linguistics exams are designed to assess metalinguistic ability—the capacity to analyze language as an object of inquiry and articulate its underlying patterns—but successful performance on such exams presupposes linguistic competence, which provides the necessary foundation for metalinguistic reasoning (Beguš et al. 2023). Metalinguistic reasoning essentially relies on an internalized grammatical system that enables speakers to generate, interpret, and intuitively evaluate linguistic forms. Thus, while linguistic exams require the explicit analysis of language, their successful completion inherently draws upon the implicit linguistic competence that underlies all language use.

Since LLMs are trained to recognize and generate recurring patterns in language data, linguistic exams offer a means to probe their higher-order metacognitive abilities in identifying and analyzing systematic patterns (Manning 2022, Millière 2024). A few studies have explored GPT's metalinguistic capacity. Beguš et al. (2023) tested GPT-4 on syntactic tasks on constituency and ambiguity, recursion, syntactic movement, and tree drawing in English and German, claiming that it may be the first LLM capable of producing coherent metalinguistic analyses. Similarly, Dąbkowski and Beguš (2023) examined linguistic recursion in syntactic phenomena, arguing that GPT-4 demonstrates cognitive abilities approaching those of humans. The present study evaluates GPT-4o on analogous syntactic questions but reaches a different conclusion: while GPT-4o demonstrates partial performance by capturing surface-level syntactic patterns, it falls short of such analytic capabilities, revealing limitations not only in metalinguistic reasoning but also in core syntactic competence.

¹ Later models, such as GPT-5 (released in August 2025), are equipped with real-time search techniques like RAG (Retrieval-Augmented Generation), which may make them less suitable for exams, as they could access online information to generate answers.

2. Syntax Exam

2.1 Test Design

The test questions were designed to evaluate GPT-4o's core syntactic knowledge, focusing on key concepts essential to human language parsing, including constituency (Q1-Q5), ambiguity (Q6-Q11), and recursion (Q12-Q15), along with c-command relations in anaphor interpretation, garden-path sentences, and center-embedding structures. The exam was created as if intended for college students with introductory syntax training, primarily using English examples but also including a few Korean examples (Q5, Q11, Q15) to test GPT-4o's ability to handle cross-linguistic parametric differences. To evaluate the model's depth of understanding beyond surface-level responses, the questions were structured in progressively complex steps: starting with definitions, advancing to simple multiple-choice or true/false questions, and culminating in tasks requiring interpretation and reasoning. As GPT-4o is a multimodal model capable of processing images as well as text, we also tested its ability to understand and generate syntactic tree diagrams for constituency and disambiguation, as in Beguš et al. (2023), without assuming any particular theoretical framework.

2.2 Test Procedure

Before the main exam, GPT-4o was pre-tested on its ability to use tree diagrams for syntactic analysis, as illustrated in (1). The results demonstrated that it could read and generate tree diagrams for simple English sentences, correctly answering the questions by identifying basic dominance and sisterhood relations.

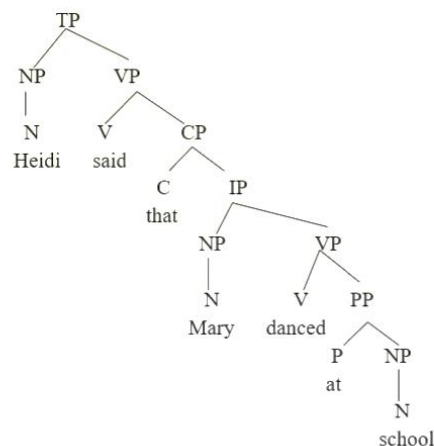
(1) Pre-test Questions

Question 1. Generate syntactic trees for the following sentences.

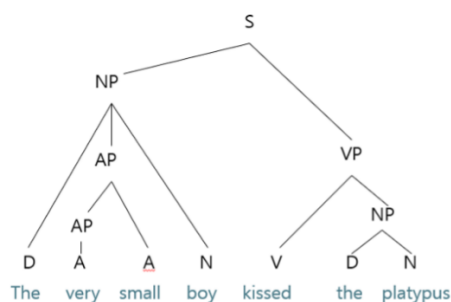
- (a) John likes Mary.
- (b) Does John eat cucumber?

Question 2. Answer the questions based on the following tree diagrams.

- (a) List all the nodes dominated by IP in the following tree.



(b) What is the sister of V in the following tree?



Then the main exam, comprising 15 syntax questions, was administered in two formats: Test1, in which each question was presented in a separate session (15 individual sessions), and Test2, where all 15 questions were delivered in a single session.² Each session began with a brief instruction to establish the model’s role and task domain: “You are a student of linguistics. Answer the following syntax question(s).”³ After both Tests were completed, we ran a few sessions of “bullying” to further probe the model’s capabilities. Here, bullying is a metaphorical term referring to persistent or pressuring behavior toward the model, such as repeatedly rephrasing or insisting on a question, or pushing the model to its limits through multiple prompts.

3. Results and Discussion

Overall, GPT-4o’s performance on syntax exam was modest compared with previous linguistics evaluations (Beguš et al. 2023) and with the model’s remarkable performances on other standardized tests such as the SAT, AP, GRE, LSAT, Bar exam, etc. (OpenAI 2023). In Test1, the model answered with detailed explanations, while in Test2 it offered only short answers. The model performed slightly better on Test1 than on Test2, contrary to our expectation that performance would improve when all questions were presented in a single session, where each question was more contextualized by preceding ones that provided additional information on the same concept. Considering only the final answers (ignoring reasoning), the model solved 11 of 15 questions in Test1 and 8 in Test2. However, when half-correct responses—those with inadequate examples, faulty reasoning, or incorrect tree parses—are excluded, the numbers drop to 5 correct in Test1 and 6 in Test2. Table 1 summarizes our syntax questions and GPT-4o’s responses.

² During Test 1, we added an additional question to further evaluate the model’s performance on Q8. This question, also included in Test2, is labeled as Q8’ in our discussion.

³ An anonymous reviewer asked why GPT-4o was assigned the role of a student rather than an expert linguist. We adopted this framing to enable comparison with college students, as many questions were drawn from actual introductory syntax exams. Our hypothesis was that if GPT-4o has basic syntactic knowledge, its answers should be comparable to those of students who have completed an introductory syntax course. One notable difference is that GPT-4o can occasionally produce correct answers even when its reasoning is flawed, whereas human students are more likely to answer incorrectly when their reasoning is unsound. For reference, Cho et al. (2025) tested GPT-4o’s phonological reasoning under varying custom instructions, reporting that expert settings such as “Professor 1” (48%) “Professor 2” (75%) and “Solver” (61%) did not necessarily outperform “Student” (70%).

Table 1. Overview of Questions and Answers

Question	Question Type	Response in Test1	Response in Test2
Q1 Define “constituent” with an example	Open-answer	Correct definition, incorrect example	Correct definition, incorrect example
Q2 Choose a non-constituent	Multiple-choice	Incorrect	Incorrect
Q3 Analyze constituents of a garden-path sentence using tree diagrams	Open-answer	Incorrect answer & tree for (A)	Incorrect answer & tree for (A)
Q4 Identify incorrect c-command relation in a tree	Multiple-choice	Correct	Incorrect
Q5 Choose the antecedent for Korean reflexive <i>caki</i>	Multiple-choice	Correct answer, faulty reasoning	Correct answer
Q6 Define “ambiguity” (structural and lexical) with examples	Open-answer	Correct definition, adequate example	Correct definition, adequate example
Q7 Identify a different type of ambiguity	Multiple-choice	Correct	Correct
Q8 Identify four interpretations of a sentence with mixed ambiguity	Open-answer	Acceptable answer, incorrect trees	Incomplete answer, acceptable trees
Q8’ Similar to Q8	Open-answer	Incorrect answer & trees	Incorrect answer & trees
Q9 Choose the correct tree for an unambiguous sentence	Multiple-choice	Correct answer, faulty reasoning	Incorrect answer
Q10 Understand an ambiguous NP	True/False	Correct answer, faulty reasoning	Incorrect answer
Q11 Disambiguate structural ambiguity of Korean sentences using trees	Open-answer	Acceptable answer, inconsistent trees	Correct answer
Q12 Define structural “recursion” with an example	Open-answer	Correct answer, correct examples	Correct answer, correct examples
Q13 Identify a different type of recursion	Multiple-choice	Correct	Correct
Q14 Understand a center-embedded sentence in a tree	Open-answer	Incorrect answer	No answer
Q15 Draw a tree diagram for a Korean center-embedded sentence and select the correct interpretation	Multiple-choice	Correct answer, incorrect tree	Correct answer, no tree

3.1 Constituency and Tree Structure

3.1.1 Constituency

When asked to define “constituent,” GPT-4o responded reasonably well in both Test1 and Test2.

User Q1: Define what a “constituent” is in linguistics. Take a sentential example and list all constituents within that sentence.

However, when listing constituents from its own example sentence, *The quick brown fox jumps over the lazy dog*, the model omitted a major constituent—PP *over the lazy dog* in Test 1 and VP *jumps over the lazy dog* in Test 2—while incorrectly including non-constituents like *The quick brown fox jumps*. It is noteworthy that this non-constituent, which the model misidentified as a constituent 3 below, could qualify as a constituent (an S/IP) if it appeared as a complete sentence without the PP. In other words, the model treated a non-constituent in the given sentence as a constituent because it can generally serve as one in other contexts.

GPT to Q1 (Test1): In linguistics, a “constituent” is a word or a group of words that functions as a single unit within a hierarchical structure. Constituents can be phrases or single words that collectively form larger structures, such as clauses or sentences. Constituents can be identified through various tests such as substitution, movement, coordination, and question-answering.

Let's take a sentential example and identify all the constituents within it.

Example Sentence: “The quick brown fox jumps over the lazy dog.”

Constituents:

1. **The quick brown fox**
2. **jumps over the lazy dog**
3. **The quick brown fox jumps**
4. **the lazy dog**

In the follow-up constituency question Q2, GPT-4o selected the incorrect answer (d) in both Test1 and Test2.

User Q2: Consider “Mary says that Susan lies to her parents a lot.” Which one of the following is not a constituent?

- (a) Mary says that Susan lies to her parents
- (b) that Susan lies to her parents a lot
- (c) lies to her parents a lot
- (d) lies to her parents

As in Q1, the model incorrectly identified (a) *Mary says that Susan lies to her parents* as a constituent. Similar to the example *The quick brown fox jumps* in the response to Q1, this string could qualify as a constituent in isolation, but not in this particular sentence in Q2 due to the presence of the adverbial *a lot*. These responses indicate that the model identifies portions of a sentence as constituents based on general grammatical patterns rather than the specific syntactic structure of the given sentence. This suggests that its grasp of constituency is not robust and is likely shaped by broad patterns of language use rather than precise structural analysis. Although it provides a reasonably accurate definition, the model appears to have only a superficial understanding of constituency, lacking a deeper grasp of the concept in relation to the internal phrase structure.

3.1.2 Garden-path sentence

Q3, adapted from Garrett et al.’s (1966) click sound test on the psychological reality of constituency, evaluates the model’s ability to identify constituents in a garden-path sentence. As shown in (A), a garden-path sentence is grammatically correct but initially misleading, requiring readers to revise their initial parse. Research on garden path sentences has shown that listeners, although taking longer to parse, eventually group words into correct constituents, indicating the active role of syntactic structure in speech perception (Frazier and Rayner 1982, Meseguer et al. 2002).

User Q3: The strings of words in (A) and (B) in the following differ in how the constituents are arranged. Draw a tree diagram for each sentence and explain the meaning of the sentence by providing a paraphrase.

- (A) In Tom’s hope of engaging Mary was idealistic.
- (B) Tom’s hope of engaging Mary was idealistic.

Despite the additional processing effort required, human speakers, drawing on their linguistic competence, can ultimately parse such sentences as (A), recognizing that unlike in (B), *Mary* is not the object of *engaging* in (A) but the subject of *was idealistic*. However, GPT-4o struggled with sentence (A) in both Test1 and Test2, as represented in Figures 1 and 2.

GPT to Q3 (Test1):

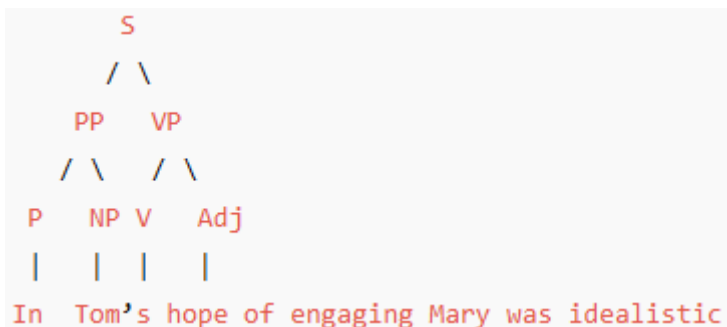


Figure 1. Tree Diagram for Sentence (A) in Q3

Paraphrase of Sentence (A)

“In the context of Tom’s hope of engaging Mary, it was idealistic.”

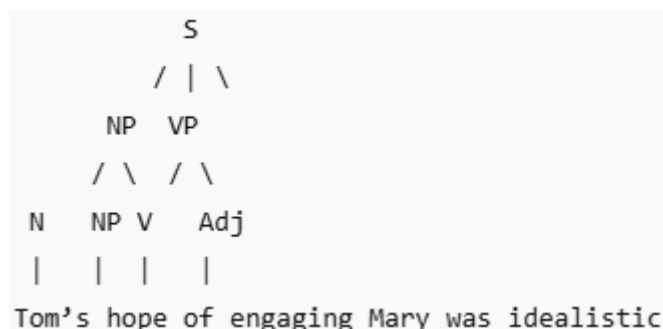


Figure 2. Tree Diagram for Sentence (B) in Q3

Paraphrase of Sentence (B)

“Tom had an idealistic hope of engaging Mary.”

Explanation of the Meaning

1. **Sentence (A):** This sentence has a prepositional phrase (“In Tom’s hope of engaging Mary”) modifying the entire clause “was idealistic.” The focus is on the entire situation or context being idealistic. It suggests that within the framework of Tom’s hope, the nature of the hope was idealistic.
2. **Sentence (B):** This sentence has the noun phrase (“Tom’s hope of engaging Mary”) as the subject of the verb “was idealistic.” The focus is on the hope itself being idealistic. It suggests that Tom had a hope, and this hope, specifically, was idealistic.

For sentence (A), GPT-4o was misled down the garden path and failed to recover, producing an absurd interpretation: “In the context of Tom’s hope of engaging Mary, it was idealistic.” Its tree diagram, shown in Figure 1, also incorrectly parsed the PP (including *Mary*) as the subject of S, a nonsensical analysis. Test2 yielded a similarly flawed reading, interpreting (A) as “It was idealistic within the context of Tom’s hope to engage Mary.” In neither Test did the model recognize *Mary* as the subject of the sentence.

To confirm the model’s inability to parse (A), we additionally conducted a “bullying” session with follow-up questions, in an attempt to guide it toward a successful parse using extra clues. However, even when explicitly informed that *Mary* may be the subject of (A), the model persisted in its original, incorrect parse, as illustrated in (2).⁴

(2) Follow-ups to Q3

- a. **User:** What is the subject of (A)?
GPT-4o: In sentence (A), “In Tom’s hope of engaging Mary was idealistic,” the subject is the prepositional phrase (PP) “In Tom’s hope of engaging Mary.”
- b. **User:** Can Mary be the subject of (A)?
GPT-4o: No, “Mary” cannot be the subject of sentence (A). . . . The subject in this sentence is the entire prepositional phrase “In Tom’s hope of engaging Mary.”

Garden-path sentences, though grammatical, are rarely used in practice due to their processing difficulty. They are “possible” but less “feasible,” and thus seldom “performed” in Hymes’s (1972) sense of communicative competence. Because such sentences are extremely rare in natural discourse, it is unlikely that the LLMs have encountered them during training, making them useful for testing the model’s competence (Chomsky, 1986). The results indicate that the model exhibits difficulty in processing sentences that are possible but rarely performed in reality, such as garden-path sentences, whereas humans can typically parse unperformed sentences through reasoning from their linguistic knowledge.

3.1.3 C-command

Next, we assessed GPT-4o’s understanding of the c-command relation in a tree diagram as represented in Figure 3. To our disappointment, the model inconsistently answered this simple question, providing the correct answer (d) in Test1, but the incorrect answer (a) in Test2.⁵

User Q4: According to the tree, which of the following is correct?

- (a) A c-commands G. (b) F c-commands B. (c) D c-commands C. (d) B c-commands G.

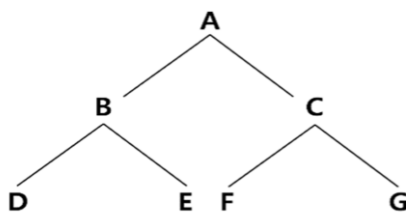


Figure 3. Tree Diagram for C-command

⁴ For space considerations, we included only the key parts of the answers for bullying sessions.

⁵ As noted by a reviewer, there are several versions of the definition of c-command, and the one we assumed follows Reinhart (1981). Regardless of which version is applied, however, only (d) satisfies all of them. Option (c) could actually be considered correct under the definition that uses the first maximal projection when A is a maximal projection and B is its intermediate projection.

Contrarily, GPT-4o performed unexpectedly well on the subsequent, more complex question about c-command in Korean anaphora.⁶ A tree structure was provided to the model as in Figure 4, and it gave the correct answer (c) in both Tests.

User Q5: Answer the following syntax question based on the tree. Korean reflexive 자기 *caki* ‘self’ is bound by the c-commanding antecedent. In the Korean example (C), which does 자기 refer back to?

(C) 선생님들께서는 철수의 친구가 자기가 최고라고 생각한다고 믿고 계신다.

Sensayngnim-tul-kkeyse-nun Cheolsu-uy chinkwu-ka caki-ka choyko-lako sayngkakhanta-ko mitko kyesinta.

Teacher-Pl-Nom(Hon)-Top Cheolsu-Gen friend-Nom self-Nom best-Comp think-Comp believe is(Hon)

‘The teachers believe that Cheolsu’s friend thinks self(*caki*) is the best.’

(a) 선생님들 ‘teachers’ (b) 철수 ‘Cheolsu’ (c) 철수의 친구 ‘Cheolsu’s friend’ (d) nobody

Figure 4. Tree Diagram for Sentence (C) in Q5

This impressive performance, however, is based on questionable reasoning, as illustrated below: “선생님들 (teachers), the subject of the main clause, does not c-command the reflexive pronoun directly within its clause.” Unlike in Q4, the model evidently recognizes c-command relationship, such that both ‘teachers’ and ‘Cheolsu’s friend’ c-command *caki* ‘self,’ although only ‘Cheolsu’s friend’ does so within its own clause.

GPT to Q5 (Test1): To determine the c-command relationship:

- The reflexive pronoun 자기 is within the embedded clause.
- Within the embedded clause, 철수의 친구 (Cheolsu's friend) c-commands the reflexive pronoun 자기.
- 선생님들 (teachers), the subject of the main clause, does not c-command the reflexive pronoun 자기 directly within its clause.

Therefore, the reflexive pronoun 자기 refers back to 철수의 친구 (Cheolsu's friend).

⁶ When Korean examples were used in the exam, neither Roman transliterations, glosses, nor English translations were provided to the model. They were added later to aid the readers’ understanding.

In reality, both (a) ‘teachers’ and (c) ‘Cheolsu’s friend’ are potential c-commanding antecedents for 자기 (*caki* ‘self’), since unlike in English, an NP outside the immediate clause such as ‘teachers’ can serve as an antecedent in Korean. ‘Teachers’ is a less suitable antecedent in (C) not because it fails to c-command the reflexive pronoun within its minimal clause but because it does not agree in number or honorific marking.⁷ If the subject were singular and non-honorific, like ‘Younghee,’ it would qualify as an antecedent. In short, GPT-4o selected ‘Cheolsu’s friend’ as the antecedent, but for a wrong reason.

To verify our interpretation of the model’s responses, we ran another “bullying” session with the following questions. As shown in (3a), the model recognized that ‘teachers’ c-commands *caki*, but consistently answered that the closest c-commanding antecedent, ‘Cheolsu’s friend,’ is the “preferred” antecedent. Even when the ‘teachers’ was replaced with ‘Younghee,’ a singular noun, as in (3b) and (3c), the model maintained that “Younghee’s being the antecedent is less likely.” Thus, although the model acknowledges that antecedent choice is a matter of preference, it never recognizes that the subject of the upper clause can be the preferred antecedent for the reflexive in the lower clause—a situation that is frequently observed, particularly when the upper-clause subject is prominent in the discourse (Kim 2000). Nevertheless, we find it noteworthy that GPT-4o produces consistent answers to the Korean reflexive binding question.

(3) Follow-ups to Q5

- a. **User:** Is 선생님들께서 also c-commanding 자기?

GPT-4o: . . . Since “선생님들께서” is at the top of the tree, it indeed dominates the entire sentence, including the node that contains “자기”. Therefore, “선생님들께서” does c-command “자기” as well. However, typically, the closest c-commanding antecedent is preferred in binding reflexives. Hence, “자기” could potentially refer to both “선생님들께서” and “철수의 친구가” but would naturally refer to the closest antecedent, which is “철수의 친구가”.

- b. **User:** In example “영희는 철수의 친구가 자기가 최고라고 생각한다고 믿고 있다”, what is the antecedent of 자기?

GPT-4o: . . . According to the principles of binding theory in Korean, the reflexive 자기 prefers the closest c-commanding antecedent within the same clause. Therefore, in this sentence, 자기 refers to the closest c-commanding antecedent, which is 철수의 친구 (Cheolsu’s friend). So, the antecedent of 자기 in this sentence is: (c) 철수의 친구 (Cheolsu’s friend).

- c. **User:** Can 영희 also be the antecedent of 자기?

GPT-4o: . . . According to the binding theory, although “영희” c-commands “자기”, the preference is typically given to the nearest c-commanding antecedent within the same clause. In this sentence, “철수의 친구” is the closest c-commanding antecedent. . . . Thus, the primary antecedent of “자기” in the given sentence is “철수의 친구”. “영희” being the antecedent is less likely according to standard binding preferences in Korean.

⁷ The reflexive *caki* (‘self’) is typically used for a plain, non-honorific antecedent and *tangshin(tul)* ‘self(Pl)’ is reserved for an honorific one, such as ‘teachers.’ Number disagreement may be a factor too as *caki* ‘self’ is singular while ‘teachers’ is plural, but speakers may vary on this judgment as plural marking is optional in Korean. Actually, *caki* can receive a distributive reading, mapping to each of the teachers, in which case number disagreement is irrelevant.

In summary, while GPT-4o correctly answered some questions on constituency, its syntactic analyses did not demonstrate genuine conceptual understanding. It provided an adequate definition but offered inadequate examples (Q1) and incorrect answers to related questions (Q2, Q3). Strikingly, it missed a simple question (Q4 in Test2), while solving a more difficult one on the same concept (Q5), yet even its correct answer relied on flawed reasoning (Q5 in Test1). Particularly telling are its misidentification of generally grammatical non-constituent clauses as constituents (Q1, Q2) and its incorrect parse of a garden-path sentence (Q3), both of which highlight its reliance on data-driven patterns rather than structural principles. The model appears to accept structures familiar from training data, while mishandling those absent, regardless of syntactic well-formedness.

3.2 Ambiguity

3.2.1 Structural ambiguity mixed with lexical ambiguity

Ambiguity, the phenomenon where a single expression yields multiple interpretations, is an inherent property of human language and a central concern in linguistic analysis. Beguš et al. (2023) reported that GPT-4 could generate distinct tree diagrams for structurally ambiguous sentences, such as *I saw an elephant with binoculars* and the less familiar *I fed my baby turtles* (p.7). Building on this finding, we tested GPT-4o to see whether it could handle cases where structural ambiguity interacts with lexical ambiguity. When asked to explain the two types of ambiguity, GPT-4o provided a textbook-like, flawless response with appropriate examples.

GPT to Q6 (Test1):

Ambiguity in Language:

Ambiguity in language occurs when a word, phrase, sentence, or even a text can have multiple meanings. This can happen at different levels of language, including syntax, semantics, and pragmatics. Ambiguity is a fundamental aspect of natural languages and can lead to misunderstandings or multiple interpretations.

Types of Ambiguity:

1. Structural Ambiguity:

Structural ambiguity, also known as syntactic ambiguity, arises when a sentence or phrase can be parsed in more than one way due to its structure. This means the same sequence of words can have multiple underlying grammatical structures.

Examples of Structural Ambiguity:

- **Sentence:** “I saw the man with the telescope.”
 - **Interpretation 1:** I used a telescope to see the man. (The phrase “with the telescope” modifies “saw.”)
 - **Interpretation 2:** The man I saw had a telescope. (The phrase “with the telescope” modifies “the man.”)
- **Sentence:** “Visiting relatives can be boring.”
 - **Interpretation 1:** The act of visiting relatives can be boring.
 - **Interpretation 2:** Relatives who are visiting can be boring.

2. Lexical Ambiguity:

Lexical ambiguity occurs when a word has more than one meaning. The ambiguity arises not from the sentence structure but from the multiple definitions or senses of a single word.

Examples of Lexical Ambiguity:

- **Word: “Bank”**
 - **Meaning 1:** A financial institution.
 - **Meaning 2:** The side of a river.
- **Word: “Bat”**
 - **Meaning 1:** A flying mammal.
 - **Meaning 2:** A piece of sports equipment used in baseball or cricket.

Key Differences:

- **Structural Ambiguity:**
 - Based on sentence structure.
 - Involves multiple ways of parsing a sentence.
 - Examples often hinge on the placement of phrases or clauses within the sentence.
- **Lexical Ambiguity:**
 - Based on word meaning.
 - Involves a single word having multiple meanings.
 - Examples often hinge on the context in which the word is used.

Understanding these types of ambiguities is crucial in fields such as linguistics, language processing, and communication, as it helps in analyzing and interpreting language more accurately.

Reassured by GPT-4o’s understanding of structural and lexical ambiguities demonstrated in its impressive response (Q6) and correct answer to the subsequent question about identifying a different type of ambiguity (Q7), we next tested its grasp of mixed ambiguities in Q8.

User Q8: Example (D) has four possible interpretations due to its lexical and structural ambiguities. List all four interpretations of (D) by providing a paraphrase and a tree diagram for each.
(D) Joan met the neighbor near the bank.

The four intended interpretations arise from the logical 2 x 2 combinations of two structural possibilities (PP *near the bank* attached to NP vs. VP) and two lexical meanings of *bank* (‘financial institution’ vs. ‘river bank’). GPT-4o, however, did not provide the full set of four. First, in Test2, Interpretations 1 and 2 represent the two structural parses (although PP attachment in parse 2 is unclear) without lexical disambiguation, while 3 and 4 demonstrate the two lexical meanings without structural disambiguation (with PP attaching to the NP). Consequently, the model missed the interpretations where the VP attachment combines with lexical ambiguity, i.e., where the meeting location is *the bank*.

GPT to Q8 (Test2):**Interpretations:**

1. Joan met the neighbor who was near the bank.
2. Joan met the neighbor at a location near the bank.
3. Joan met the neighbor who was close to a financial institution.
4. Joan met the neighbor near a riverbank.

Tree Diagrams:

1. [Joan [met [the neighbor [near the bank]]]]

2. [[Joan met [the neighbor]] [near the bank]]
3. [Joan [met [the neighbor [near [the financial institution]]]]]]
4. [Joan [met [the neighbor [near [the riverbank]]]]]]

In Test1, GPT-4o’s performance was incoherent. While the Paraphrases appeared to capture the four ambiguities, the Interpretations did not align with them, particularly in (a) and (b). If the NP, not the VP, is modified (as suggested in the Paraphrases (a) and (b)), the PP *near the bank* does not necessarily mean the “meeting location.”

GPT to Q8 (Test1):

a. **Interpretation 1:** The meeting location is near a financial institution.

Paraphrase: Joan met the neighbor who was near a financial institution.

Tree Diagram:

```
(S
  (NP (N Joan))
  (VP
    (V met)
    (NP (Det the) (N neighbor))
    (PP (P near) (NP (Det the) (N bank))))))
```

b. **Interpretation 2:** The meeting location is near a river bank

Paraphrase: Joan met the neighbor who was near a river bank.

c. **Interpretation 3:** The meeting happened near a financial institution

Paraphrase: Near the financial institution, Joan met the neighbor.

Tree Diagram:

```
(S
  (PP (P near) (NP (Det the) (N bank)))
  (S
    (NP (N Joan))
    (VP (V met) (NP (Det the) (N neighbor))))))
```

d. **Interpretation 4:** The meeting happened near a river bank

Paraphrase: Near the river bank, Joan met the neighbor.

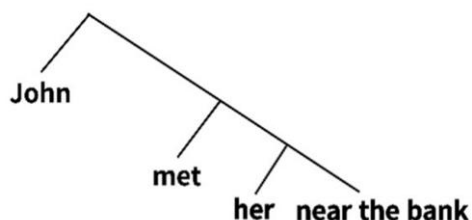
Moreover, GPT-4o produced unmatching tree structures for its Interpretations or Paraphrases. It gave identical parses for (a) and (b) and another for (c) and (d) (only one tree per pair was displayed above due to space limitations), all of which basically placed the PP outside the NP. For (a) and (b), the PP was inside the VP, aligning with the Interpretations but not the Paraphrases. For (c) and (d), the PP was adjoined to S clause-initially, which indicates non-attachment to the NP but still mislocates the PP. In short, the trees do not align with the given interpretations or paraphrases, suggesting that GPT-4o fails to produce tree structures consistent with its own disambiguation analysis.

Questioning the model’s structural understanding, we then tested a version where the object NP *the neighbor* was replaced with the pronoun *her*, thereby eliminating structural ambiguity. The model was given two trees to choose from, as shown in Figure 5. In this case, only structure (b), where the PP attaches to the VP, remains a viable parse.

User Q9: If you substitute *the neighbor* in (D) with *her* as in (E), the sentence becomes only two-way ambiguous. Which of the following tree diagrams describes the structure of (E)?

(E) Joan met her near the bank.

(a)



(b)

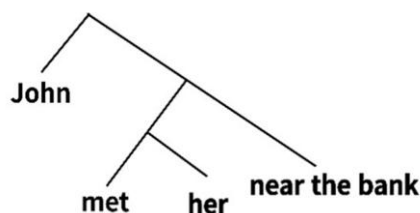


Figure 5. Tree Diagrams for Two-way Ambiguity

In Test1, GPT-4o selected the correct answer (b), but its reasoning was questionable, as its explanation of tree (a) is practically identical to that of (b).

GPT to Q9 (Test1):

Tree (a): Prepositional phrases (near the bank) that modify the verb are also connected directly to the verb, often through a prepositional phrase (PP) node.

Tree (b): “near the bank” is a prepositional phrase functioning as an adverbial modifier describing the location of the action.

In Test2, GPT-4o selected the incorrect answer (a). The model failed to consistently yield correct answers across both Tests. Even when it answered correctly in Test1, its response did not appear to be based upon accurate understanding of the structural difference, offering indistinct parses.

To assess whether this disambiguation failure stemmed from the model’s lack of understanding of tree diagrams or ambiguity itself, we conducted an additional test using a more familiar example (D’), similar to that examined by Beguš et al. (2023) (*I saw an elephant with binoculars*) and to the model’s own example for Q6 (*I saw the man with the telescope*).

User Q8’: Example (D’) has multiple interpretations due to its lexical and structural ambiguities. List all possible interpretations of (D’) by providing a paraphrase and a tree diagram for each.

(D’) Joan drew a picture with a mouse.

In both Test1 and Test2, the model initially provided only two interpretations, which likely reflect more common real-world scenarios. The first “instrumental” interpretation attached the PP *with a mouse* to the VP, lexically reading *mouse* as a ‘computer mouse.’ The second “accompaniment” interpretation (although the title does not match the interpretation) attached the PP to the NP, reading *mouse* as an ‘animal mouse.’

GPT to Q8’ (Test1):

Example (D’) “Joan drew a picture with a mouse” has two primary interpretations due to its lexical and

structural ambiguities:

1. **Instrumental Interpretation:** Joan used a mouse (computer mouse) to draw a picture.
2. **Accompaniment Interpretation:** Joan drew a picture that included a mouse (the animal) in it.

However, in both Tests, GPT-4o failed to provide acceptable tree structures, particularly for the first instrumental-PP interpretation. The first tree in Test1 was unreasonable (V takes PP as its object), and the first parse in Test2 was incorrect for the interpretation (the first and second parses in Test2 are practically identical, corresponding to the NP-modifying interpretation). Compare the attachments in Figures 6 and 7 generated in Test 1 and the brackets provided in Test 2.

GPT to Q8’:

[Test1]

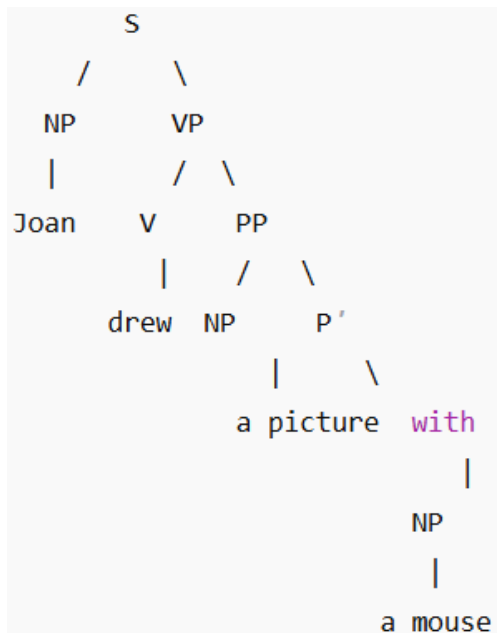


Figure 6. First Tree Diagram for Sentence Q8’

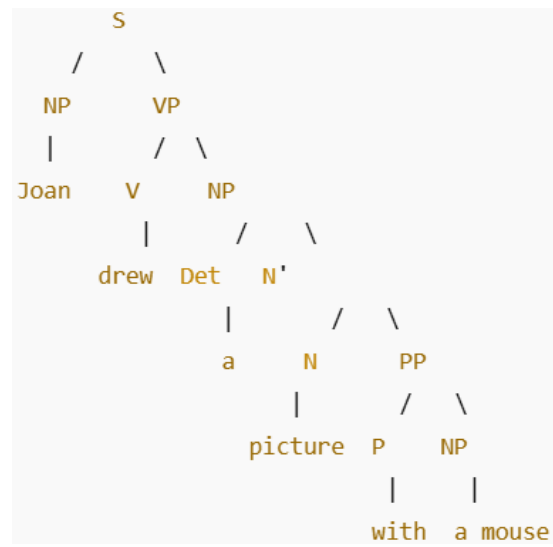


Figure 7. Second Tree Diagram for Sentence Q8’

[Test2]

1. [Joan [drew [a picture [with [a mouse]]]]]
2. [Joan [drew [a picture [that includes [a mouse]]]]]

When prompted further for additional interpretations, GPT-4o gave the following four, with 1 and 2 repeating the previous ones. Interpretation 3, though somewhat odd, could be seen as derived from the VP-attached PP parse (similar to 1) with *mouse* interpreted as an ‘animal mouse.’ The missing interpretation is the NP-attached PP with *mouse* meaning a ‘computer mouse,’ i.e., ‘Joan drew a picture that includes a computer mouse.’ Despite its conceptual plausibility, this interpretation never appeared. Instead, GPT-4o offered an absurd “combined” interpretation in 4, which was both structurally and semantically impossible, requiring *mouse* to mean a ‘computer mouse’ and an ‘animal mouse’ simultaneously.

GPT to Q8' (Test1):

In summary, the sentence “Joan drew a picture with a mouse” can be interpreted in the following ways:

1. Joan used a mouse (instrumental) to draw a picture.
2. Joan drew a picture that includes a mouse (accompaniment).
3. A mouse drew a picture, and Joan assisted (agentive, though less plausible).
4. Joan used a mouse to draw a picture that includes a mouse (combined interpretation)

Why does the model struggle so much to generate the missing, logically plausible reading of ‘drawing a picture that includes a computer mouse’? A quick image search for “picture drawing that includes a mouse” offers a clue. As seen in Figure 8, real-world data rarely feature ‘a picture that includes a computer mouse,’ whereas depictions of an anthropomorphized animal mouse potentially capable of drawing a picture are much more common. This suggests that GPT-4o more likely generates interpretations that align with more frequent real-world scenarios, rather than logically reasoning about all possibilities, including less probable ones.

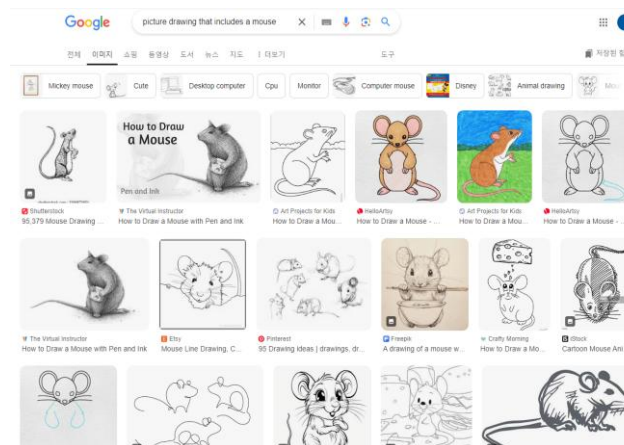


Figure 8. Picture Drawing that Includes a Mouse

In fact, we “bullied” the model by prompting it in various ways to elicit the logically possible fourth interpretations, but all attempts failed. Finally, when we directly asked whether the missing interpretation that ‘John drew a picture that includes a computer mouse’ was a possible one, the model at last said yes.

(4) Follow-ups to Q8'

User: Is it possible that John drew a picture that includes a computer mouse?

GPT-4o: Yes, it is possible that “John drew a picture with a mouse” could mean that John drew a picture that includes an image of a computer mouse. This would add another interpretation to the sentence.

Finally, the tree diagrams that accompanied the initial four interpretations given by the model, shown in Figures 9 to 12, were no better than those for the two interpretations shown in Figures 6 and 7.

GPT to Q8' (Test1):

- 1. Instrumental Interpretation: (Paraphrase)** Joan used a mouse to draw a picture.
- 2. Accompaniment Interpretation: (Paraphrase)** Joan drew a picture that includes a mouse.
- 3. Agentive Interpretation: (Paraphrase)** A mouse drew a picture, and Joan was involved or assisted in some

way.

4. Combined Interpretation: (Paraphrase) Joan used a mouse to draw a picture that includes a mouse.

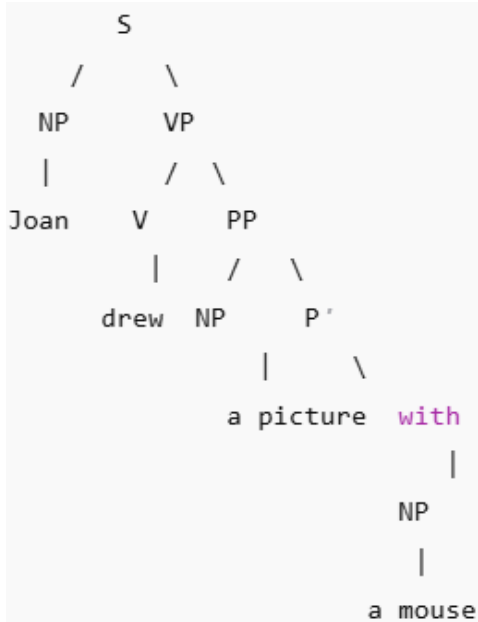


Figure 9. Tree Diagram for the First Interpretation

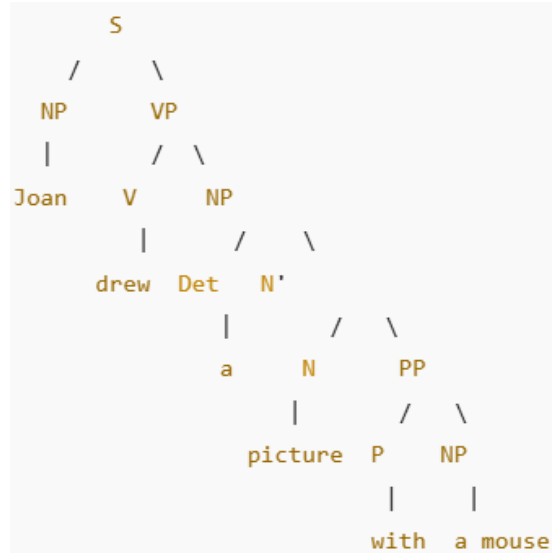


Figure 10. Tree Diagram for the Second Interpretation

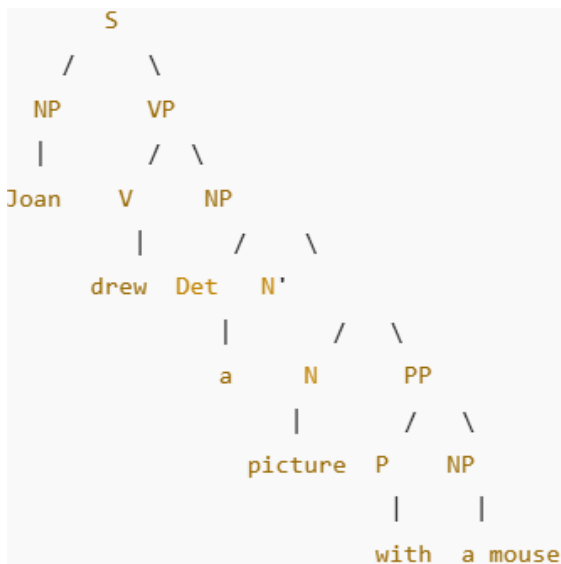


Figure 11. Tree diagram for the Third interpretation

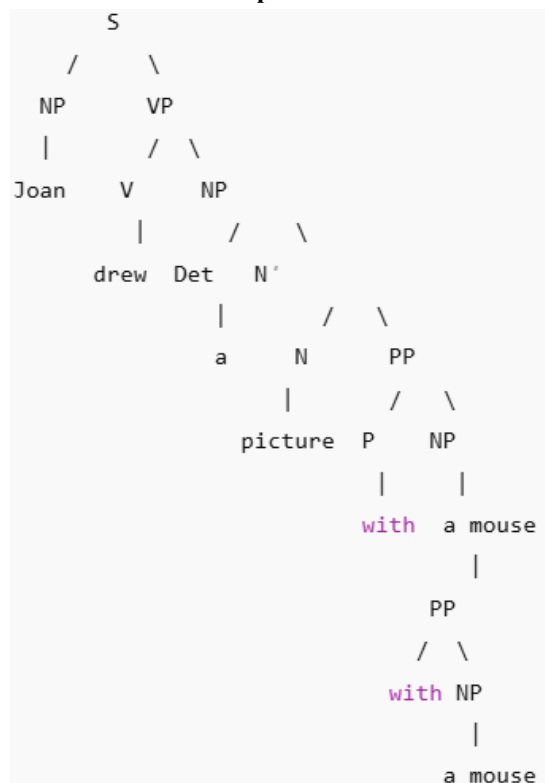


Figure 12. Tree Diagram for the Fourth Interpretation

The model provided the same incorrect tree for interpretation 1, used a single tree for the different structural interpretations 2 and 3, and generated an unintelligible tree for interpretation 4. These results do not align with those of Beguš et al. (2023), who argued that GPT-4 disambiguated the structurally ambiguous sentences such as *I saw an elephant with binoculars* and *I fed my baby turtles* with appropriate tree diagrams. Setting the lexical ambiguity aside, GPT-4o should at least have produced distinct trees to reflect the structural ambiguity. GPT-4o's poor performance on tree diagrams is therefore unexpectedly disappointing, given that it is a more advanced multimodal model than GPT-4, capable of processing images such as tree diagrams, and that it successfully passed the pretest for tree generation and interpretation. One possible explanation is that Beguš et al. (2023) prompted GPT-4 to “do a theoretical syntactic analysis of the sentence in LaTeX using the forest package and to use the X-bar theory.” The model then produced tree structures in LaTeX code, which were subsequently rendered as tree diagrams based on the specific X-bar framework. In other words, the model was operating under tightly constrained instructions within the LaTeX coding environment, which is widely used for drawing linguistic trees. It may therefore have been better trained in this particular task. In general, language models are known to perform better on coding tasks than on natural language, due to the more constrained and predictable nature of coding contexts. In other words, GPT-4 might not have truly understood tree structures but was instead more accustomed to drawing trees in that specific context. Given no comparable prompt in our test and thus deprived of such a specifically trained environment, GPT-4o clearly revealed its lack of understanding of tree structures.

In summary, despite providing seemingly “perfect” conceptual definitions for ambiguity (Q6) and correctly identifying its different types (Q7), GPT-4o failed to identify all four interpretations arising from structural and lexical ambiguities (Q8 and Q8') and struggled to consistently recognize the absence of ambiguity with a pronoun (Q9). While its inability to generate accurate tree structures, as also seen in the previous subsection, further underscores this limitation, it is evident that the model cannot reliably perform syntactic analyses of disambiguation, independent of its ability to use tree diagrams as a meta-linguistic tool.

3.2.2 Ambiguity in NP

The next question addresses ambiguity within noun phrases. Donyms such as *English* and *Korean* can refer to either nationality/ethnicity or language. Interestingly, however, while *Korean* in *the Korean teacher* is ambiguous between nationality and language, in *the Japanese Korean teacher*, *Korean* unambiguously refers to the language. Therefore, the answer to Q10 is false.

User Q10: Is the following statement true or false? Explain why.

The interpretations of “Korean” in “the Korean teacher” and “the Japanese Korean teacher” are unambiguous but different in each case: “Korean” in “the Korean teacher” means the person with Korean nationality while it means the Korean language in “the Japanese Korean teacher.”

However, GPT-4o offered inconsistent answers between Test1 and Test2 again: the correct answer “false” in Test1 but the incorrect “true” in Test2. Moreover, the explanation for the “correct” answer in Test1 was flawed. The model claimed that the *Korean* in the *Japanese Korean teacher* was also ambiguous, including the implausible and unacceptable reading (c), in which the *Japanese* in the *Japanese Korean teacher* is understood as the language and *Korean* as nationality/ethnicity. This interpretation is impossible because the closer adjective to the noun *teacher* must be interpreted as its complement—that is, the object of teaching, hence the language.

GPT to Q10 (Test1):1. **The Korean teacher:**

- This phrase can mean:

- a. A teacher who is of Korean nationality or ethnicity.
- b. A teacher who teaches the Korean language.

Without additional context, both interpretations are possible.

2. **The Japanese Korean teacher:**

- This phrase is also ambiguous and can mean:

- a. A teacher who is ethnically Korean but has Japanese nationality.
- b. A teacher who is Japanese and teaches the Korean language.
- c. A teacher who is Korean and teaches Japanese.

Again, the interpretation depends on the context provided.

We ran a bullying session with the reverse example *the Korean Japanese teacher* to reconfirm the model's interpretation of (c), as illustrated in (5). The model again listed the impossible interpretation 4 in (5a), in which *Korean* is interpreted as the language. When pressed further, the model eventually conceded, albeit awkwardly, that “the Korean Japanese teacher would not typically refer to a teacher of the Korean language, even in untypical situations,” as shown in (5c) and (5d).

(5) Follow-up to Q10

- a. **User:** What are possible interpretations of “the Korean Japanese teacher”?

GPT-4o:1. **A Japanese teacher of Korean ethnicity or nationality:**

- This could mean a teacher who is ethnically Korean but has Japanese nationality or is residing in Japan.

2. **A teacher who is both Korean and Japanese:**

- This interpretation could imply a person with dual heritage or dual nationality, identifying as both Korean and Japanese.

3. **A teacher of the Japanese language who is Korean:**

- This could mean a teacher who is ethnically or nationally Korean but teaches the Japanese language.

4. **A teacher of the Korean language who is Japanese:**

- This could mean a teacher who is ethnically or nationally Japanese but teaches the Korean language.

5. **A teacher who is Korean and teaches Japanese:**

- This interpretation might focus on the profession, indicating a teacher who is Korean and whose subject of instruction is the Japanese language.

- b. **User:** What's the difference between 3 and 5?

GPT-4o: . . . Upon closer examination, interpretations 3 and 5 are essentially describing the same situation.

- c. **User:** Are you sure that “the Korean Japanese teacher” can refer to a teacher of the Korean language?

GPT4-o: Upon closer analysis, “the Korean Japanese teacher” would not typically refer to a teacher of the Korean language. . .

d. **User:** Do you mean that “the Korean Japanese teacher” could refer to a teacher of the Korean language in an untypical situation?

GPT-4o: No, “the Korean Japanese teacher” would not typically refer to a teacher of the Korean language, even in untypical situations. . .

Another notable point in GPT-4o’s response to Q10 is that the model provided a plausible yet unexpected reading (2a), interpreting *Japanese Korean* as a (hyphenated) compound referring to an ‘ethnic Korean with Japanese nationality,’ analogous to *American Jewish* vs. *Polish Jewish*. This interpretation warrants closer attention. In actual practice, however, *Japanese Korean* is more likely to be interpreted as ‘ethnically Japanese but has Korean nationality,’ analogous to *Japanese American* vs. *Chinese American*, since nationality tends to carry greater significance than ethnicity in modern contexts and therefore is typically positioned later in the phrase structure.

Interestingly, in the follow-up question in (5), the model interpreted the reversed phrase *Korean Japanese* in the same way—again as ‘ethnic Korean with Japanese nationality,’ as illustrated as 1. In other words, the model treats both *Japanese Korean* and *Korean Japanese* as meaning ‘ethnic Korean with Japanese nationality,’ while the opposite reading—‘ethnic Japanese with Korean nationality’—is conspicuously absent. This asymmetry seems to mirror socio-historical reality rather than grammatical possibility. In fact, individuals of Japanese ethnicity with Korean nationality are rare, whereas those of Korean ethnicity with Japanese nationality are far more common, often referred to as *Zainichi*, a term rooted in the historical relationship between the two nations.

The model’s biased nationality-ethnicity interpretation—favoring the less common syntactic pattern rather than the more common ethnicity-nationality reading—likely stems from its data-driven learning. As in earlier examples, GPT-4o appears to disproportionately favor frequency-based patterns over analogical reasoning based on the morphological or syntactic structure of the NP. This recurring reality-based bias raises doubts about whether the model’s apparent syntactic analyses reflect genuine reasoning grounded in linguistic knowledge, or merely the statistical patterns embedded in its training data.

3.2.3 Ambiguity in Korean

Finally, we assessed the model’s ability to analyze Korean ambiguous sentences.

User Q11: Are the following Korean sentences ambiguous? Explain the different interpretations with tree diagrams.

(F) 영희가 매일 도서관에 가는 친구에게 전화한다.

Yenghee-ka mayil tosekwan-ey ka-nun chinkwu-eykey cenhwaha-n-ta

Yenghee-Nom everyday library-to go-Mod friend-to call-Prs-Dcl

‘Younghee calls her friend who goes to the library every day.’

(G) 세명의 간호사와 의사가 응급실에 들어왔다.

Seymyeong-uy kanhosa-wa uysa-ka ungupsil-ey tulewa-ss-ta

Three-Gen nurse-and doctor-Nom emergency room-to enter-Pst-Dcl

‘Three nurse(s) and doctor(s) entered the emergency room.’

For example (F), the model correctly identified the ambiguity of what 매일 *mayil* ‘every day’ modifies (although

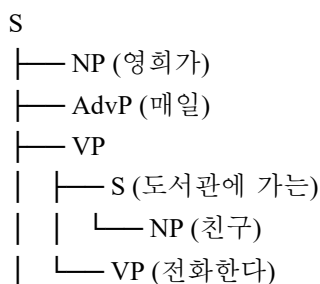
Interpretation 2 remains questionable) in both Test1 and Test2. However, in the tree diagrams, it failed to place the AdvP ‘every day’ inside the relative clause S for Interpretation 1, not to mention other inadequacies. The model provided no trees at all in Test2.

GPT to Q11F (Test1):

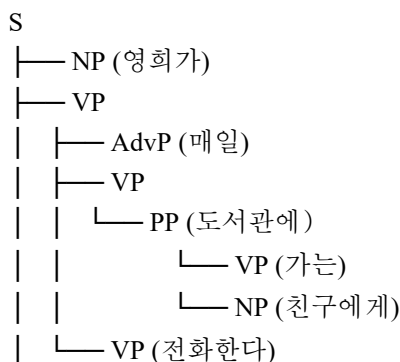
(F) 영희가 매일 도서관에 가는 친구에게 전화한다.

This sentence can be interpreted in two different ways due to the attachment ambiguity of the phrase “매일 도서관에 가는” (who goes to the library every day).

a. Interpretation 1: Younghee calls a friend who goes to the library every day.



b. Interpretation 2: Younghee calls a friend every day to go to the library.



For example (G), the model generated two readings based on the position of the quantifier 세명의 (*semyeong-uy* ‘three’) and provided disambiguating trees, albeit with some imperfections: Interpretation 1 corresponded to ‘three nurses and one doctor’ and Interpretation 2 to ‘three nurses and three doctors.’ In Korean, plural marking is optional, so a bare noun can in principle be interpreted as singular or plural.

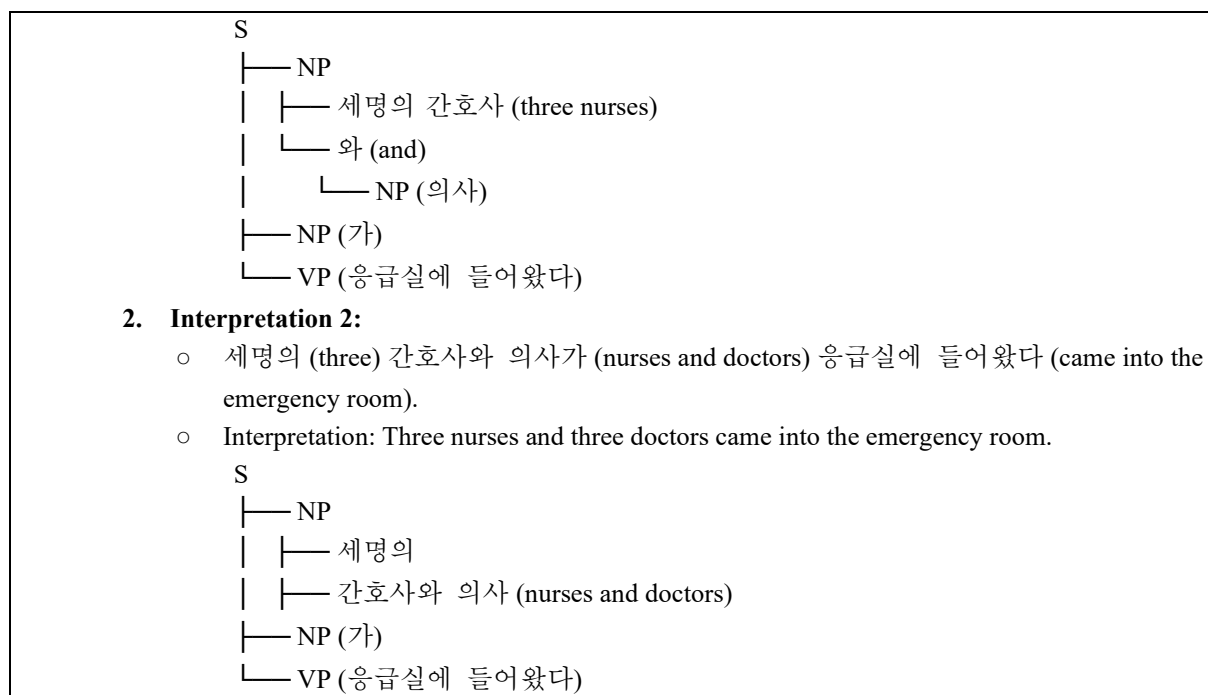
GPT to Q11G (Test1):

(G) 세명의 간호사와 의사가 응급실에 들어왔다.

This sentence can be interpreted in two different ways due to the coordination ambiguity of the phrase “세명의 간호사와 의사” (three nurses and a doctor).

1. **Interpretation 1:**

- 세명의 (three) 간호사와 (nurses and) 의사가 (doctor) 응급실에 들어왔다 (came into the emergency room).
- Interpretation: Three nurses and one doctor came into the emergency room.



For Interpretation 2, the model applied a distributive reading of ‘three’ across ‘nurses and doctors,’ resulting in six individuals, although a non-distributive reading—three individuals in total—is also possible. Interestingly, in Test2, the model produced the non-distributive reading while the distributive reading was absent, as shown below.⁸

GPT to Q11G (Test2):

(b) 세명의 간호사와 의사가 응급실에 들어왔다.

1. (세 명의 간호사)와 (의사) (entered the emergency room). (Three nurses and a doctor entered the emergency room.)
2. (세 명의 (간호사와 의사)) (entered the emergency room). (Three individuals, including nurses and a doctor, entered the emergency room.)

In this analysis of Korean examples, GPT-4o identified ambiguity reasonably well, though not flawlessly, but showed inconsistent performance in generating disambiguating trees—unacceptable for (F) but somewhat acceptable for (G), if minor tree details are disregarded.

In summary, this subsection reveals patterns consistent with those observed earlier: (i) near-perfect definitions of key concepts; (ii) seemingly correct answers often supported by flawed reasoning; (iii) incorrect answers stemming from data-driven biases; and (iv) difficulties in accurately interpreting or generating tree diagrams. Interestingly, the model’s handling of Korean examples does not appear markedly weaker overall.

⁸ In this non-distributional reading, three individuals, including a nurse and doctors, is also a possible interpretation, but the model apparently shows a tendency to associate the plurality with the closer noun *nurse* to the quantifier.

3.3 Recursion and center embedding

The final test area concerns recursion, another unique property of human language that allows for multiple embeddings of phrases and clauses, thereby allowing the creation of an infinite number of sentences in complex hierarchical structures. In their evaluation of language models' capacities, Dabkowski and Beguš (2023) reported that GPT-4 could generate, comprehend, and analyze recursive structures, capable of constructing corresponding tree diagrams. Recursion can produce center-embedded sentences, in which clauses are nested within another. Like the garden-path sentences discussed in section 3.1, such constructions are grammatical but difficult to parse, often requiring careful rereading to fully grasp their meaning. Precisely because of this processing difficulty, they are rarely used in practice, yet they serve as a valuable diagnostic for assessing the model's underlying linguistic competence.

When asked to define "recursion" in Q12 and identify its types in Q13, GPT-4o performed excellently, just as in previous subsections, consistent with the findings of Dabkowski and Beguš' (2023). Center-embedding was then addressed in Q14. Beguš et al. (2023) reported that GPT-4 "performs remarkably well" on center-embedding tasks and produced a reasonable tree diagram for sentences like *The mouse that the cat that the dog painted taught sang* (p. 11). We presented GPT-4o with a structurally similar center-embedded sentence on a tree diagram to test its ability to interpret the hierarchical structure to derive the correct reading. To avoid any theory-internal assumptions, node labels were deliberately omitted, as shown in Figure 13.

User Q14: The following is a tree diagram of an English sentence. Describe what each animal did according to the tree structure.

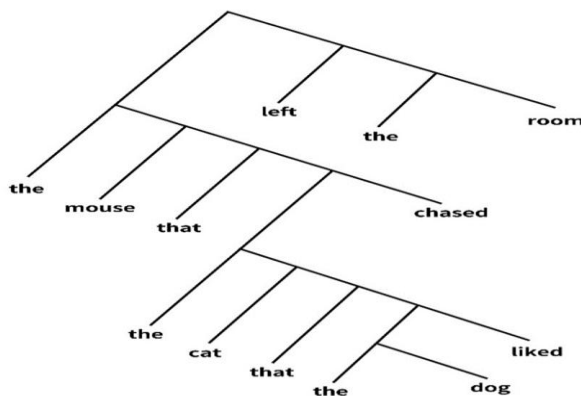


Figure 13. Tree Diagram for Q14

Contrary to expectations, GPT-4o completely failed to understand the sentence *The mouse that the cat that the dog liked chased left the room*. The model mismatched the two embedded subjects with the wrong verbs, incorrectly interpreting the sentence as meaning 'The mouse that chased the cat that liked the dog left the room,' as illustrated below. In Test2, the model failed to read the tree altogether and provided no response.

GPT to Q14 (Test1):

- a. The dog:** The dog is liked by the cat. This is derived from the subtree “the cat that liked the dog.”
- b. The cat:** The cat that liked the dog is chased by the mouse. This is derived from the subtree “the mouse that chased the cat that liked the dog.”
- c. The mouse:** The mouse that chased the cat that liked the dog left the room. This is derived from the full sentence represented in the tree, which starts with “the mouse that chased the cat that liked the dog left the room.”

Obviously, GPT-4o incorrectly disentangled the given center-embedded sentence in Q14 into a non-linear, non-center-embedded one. Apparently, the error arises from the model’s way of processing the tree in Figure 13: the model appears to process the words in the tree in a mixed top-down and left-to-right fashion without recognizing the hierarchical clause structures. For instance, after reading *the mouse that* at the top left of the tree, it skipped directly to the verb *chased*—which is neither the next linear word nor the matching verb—apparently because *chased* was positioned next at the same “height” in the tree. After *chased*, the model appears to have dropped down to the next level to read the word strings *the cat that* and again jumped to the verb *liked* at the same height, then to *the dog* at the next level down, and finally to the remaining part *left the room* at the top right. Simply put, the model just read the word strings layer by layer from the top left to down right, disregarding the branching structures of the tree. This peculiar reading pattern reveals a fundamental failure to grasp hierarchical dominance and precedence relations between nodes, leading to the inevitable conclusion that GPT-4o does not comprehend tree diagrams.

To determine whether the failure stemmed from a lack of understanding of tree structures or of the center-embedding itself, we proceeded with follow-up questions presented in plain text without tree diagrams, as in (6a). However, the model produced the same incorrect interpretation—‘The dog is liked by the cat’ and ‘The cat is chased by the mouse’—as it had when given the tree structure in Q14.

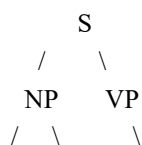
(6) Follow-ups to Q14

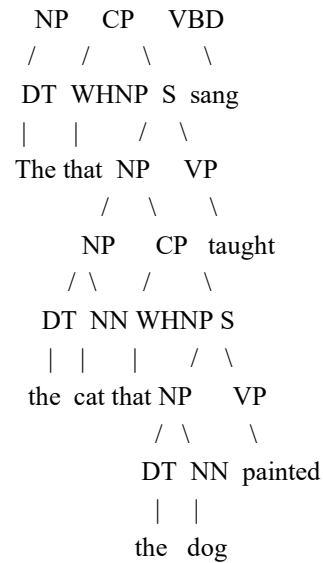
- a. **User:** Now consider the example “The mouse that the cat that the dog liked chased left the room.” Describe what each animal did in this sentence.

GPT-4o:

In the sentence “The mouse that the cat that the dog liked chased left the room” we have a nested relative clause structure. Here's the breakdown of what each animal did:

1. **The dog:** The dog is liked by the cat. This is derived from the phrase “the cat that the dog liked.”
 2. **The cat:** The cat that the dog liked is chased by the mouse. This is derived from the phrase “the mouse that the cat that the dog liked chased.”
 3. **The mouse:** The mouse that the cat that the dog liked chased left the room. This is derived from the entire sentence.
- b. **User:** Draw a syntax tree structure of “The mouse that the cat that the dog painted taught sang.”

Chat-4o:



- c. **User:** Describe what each animal did according to the above tree structure.

ChatGPT:

Based on the tree structure for the sentence “The mouse that the cat that the dog painted taught sang,” here's what each animal did:

The dog: The dog was painted by someone. This is derived from the clause “the dog painted.”

The cat: The cat, which the dog painted, taught someone or something. This is derived from the clause “the cat that the dog painted taught.”

The mouse: The mouse, which the cat that the dog painted taught, sang. This is derived from the main clause “The mouse that the cat that the dog painted taught sang.”

Even when subsequently asked about the same sentence as was tested by Beguš et al. (2023), *The mouse that the cat that the dog painted taught sang*, in (6b) and (6c), the model yielded an odd reading in which ‘The dog was painted by someone’ and ‘The cat taught someone or something,’ despite producing a plausible tree structure when the specific details were ignored. Therefore, we cannot help but conclude that GPT-4o does not truly understand center-embedding, regardless of its apparent ability to handle tree structures. Contrary to Beguš et al.’s claim (2023), the model is at best inconsistent in its performance on center-embedding tasks.

By contrast, when presented with a center-embedded sentence (H) in Korean in Q15, GPT-4o successfully parsed it and selected the correct answer (d) in both Tests. As a head-final SOV language, Korean constructs center-embedding through multiple layers of complement clauses: [S [S [S O V] V] V].⁹

User Q15: Draw a tree diagram for the following Korean sentence (H). Which of the following is correct about this sentence?

(H) 철수는 친구가 매일 영희가 도서관에서 노래를 부르는 것을 보았다고 말했다.

Chulsu-neun chingu-ga mayil Younghee-ga tosegwan-eyse nolay-lul bureuneun geos-ul boassdago malhaetta

⁹ The Roman transliterations, which are added later, are not consistent among Korean examples to match the ones used inconsistently by GPT-4o in its responses.

Chulsu-Top friend-Nom everyday Younghee-Nom library-in song-Acc sing saw said

‘Chulsu said that the friend saw Younghee sang a song at the library every day.’

- (a) 철수가 영희를 불렀다. ‘Chulsu called Younghee.’
- (b) 철수가 노래를 불렀다. ‘Chulsu sang a sing.’
- (c) 영희가 도서관에서 말했다. ‘Younghee said at the library.’
- (d) 친구가 영희를 보았다. ‘The friend saw Younghee.’

This was indeed an unexpected result. How could a model that failed to parse center-embedding in English succeed with a structurally analogous sentence in Korean? Both cases involve center-embedding with the same depth and subject-verb linear order, yet the model struggled with the English tree, while successfully interpreting the Korean sentence. One plausible explanation may be that the Korean example involves a regular *that*-clause type of embedding, while the English examples contain relative clause embedding, which is generally more difficult to process. Alternatively, the model may rely on surface heuristics—such as matching subject-verb pairs from the inside out in Korean. Either way, we can offer no better explanation than that the model is inconsistent and lacks solid linguistic knowledge. This discrepancy further raises doubts about GPT-4o’s ability to understand hierarchical structures in a way that reflects genuine structural comprehension.

GPT to Q15 (Test1):

To construct the tree diagram, we need to recognize the hierarchical structure of the sentence. The main clause is “철수는 ... 말했다” (Chulsoo-neun ... malhaetta). The embedded clause is “친구가 ... 보았다고” (chingu-ga ... boassdago), and within this embedded clause is another embedded clause “영희가 ... 부르는 것” (Yeonghee-ga ... bureuneun geos).

So, the correct answer is: (d) 친구가 영희를 보았다. (The friend saw Yeonghee.)

In summary, while defining “recursion” accurately with examples (Q12) and correctly identifying its types (Q13), GPT-4o failed to parse a typical center-embedded sentence in English (Q14), contradicting Beguš et al.’s (2023) claim. Unlike the ambiguity questions seen in section in 3.2, this failure is not due to difficulties with tree diagrams, as is evident in the follow-up probing shown in (6). By contrast, it succeeded with a structurally similar Korean sentence (Q15), revealing a discrepancy in its syntactic processing. Overall, the model did not exhibit greater difficulty with Korean examples than English. These findings suggest that GPT-4o’s metalinguistic analysis of recursion and center-embedding is not robust, and its apparent success does not reflect genuine linguistic and metalinguistic competence in handling complex syntactic structures.

4. Conclusion

The syntax exam results indicate that while demonstrating impressive surface-level performance in some syntactic tasks, GPT-4o fails to exhibit genuine linguistic or metalinguistic competence. Although the model often provides accurate definitions and seemingly correct outputs, these are frequently based on flawed reasoning or heuristics rather than a true understanding of syntactic principles. This limitation is evident across tested domains, such as constituency, ambiguity, and recursion. Moreover, inconsistencies in performance—both across test sessions and within individual responses—highlight its lack of systematic syntactic reasoning. These findings

underscore that GPT-4o relies primarily on statistical pattern recognition rather than actual linguistic or metalinguistic competence.

Behavioral linguistics tests—such as evaluating grammaticality or performing syntactic operations like agreement, movement, or embedding—assess only language performance (Haider 2023), making it difficult to determine whether correct outputs arise from distributional knowledge or understanding of linguistic structure. In contrast, prompting a model to analyze language structure in linguistic terms provides deeper insight into its ability to access hierarchical linguistic structures to guide its behavioral outputs (Beguš et al. 2023). Learning language solely through data restricts the model to what has already been “performed,” biasing it toward the “feasible” without fully learning the “possible” or true linguistic competence (Hymes 1972). Tests of constituency in garden-path sentences, ambiguity, and center-embedding are particularly revealing in evaluating a model’s competence or its grasp of the “possible,” i.e., the core linguistic knowledge. Contrary to Beguš et al.’s (2023) claim that LLMs, particularly GPT-4, can generate “coherent and valid formal analyses of linguistic data” and demonstrate potential for “analyses of metalinguistic abilities,” our findings suggest otherwise. Setting aside its difficulty with tree diagrams, GPT-4o’s incoherent performance in all three areas demonstrates its lack of competence. This study, however, has a limitation of reliance on English examples, with only a few Korean examples included, which constrained a fully parallel comparison of the model’s performance across the two languages. Future research should pursue cross-linguistic investigation more systematically.

Human linguists develop analytical tools foundational to linguistic science by reasoning about language structure grounded in their native-language knowledge—that is, their mental grammar. Whether large language models (LLMs) can develop a comparable capacity to analyze their own grammar through linguistic reasoning, however, remains uncertain. Our findings suggest that the LLM’s behavioral performance is not guided by linguistic structure and does not reflect their linguistic or metalinguistic competence. Ultimately, this study supports skepticism regarding LLM’s capacity for genuine linguistic competence, demonstrating its lack of capacity for analytical reasoning about language structures akin to that of human linguists. Addressing these shortcomings may require future advancements in LLM architecture, as current models remain bound by training data and lack the fundamental grammatical intuitions intrinsic to human cognition.

References

- Beguš, G., M. Dąbkowski and R. Rhodes. 2023. Large linguistic models: Analyzing theoretical linguistic abilities of LLMs. *arXiv preprint* arXiv:2305.00948.
- Bender, E. M., T. Gebru, A. McMillan-Major and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623
- Chang, T. A. and B. K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics* 50(1), 293-350.
- Cho, H., S. Park, S. Song and E. E. Oh. 2025. Investigating ChatGPT’s phonology problem-solving abilities through reasoning with varying custom instructions. *Linguistic Research* 42(1), 53-93.
- Chomsky, N. 2000. Knowledge of language: Its nature, origin, and use. In R. J. Stainton, ed., *Perspectives in the Philosophy of Language: A Concise Anthology*, 3-44.
- Dąbkowski, M. and G. Beguš. 2023. Large language models and (non-) linguistic recursion. *arXiv preprint* arXiv:2306.07195.

- Frazier, L. and K. Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology* 14(2), 178-210.
- Garrett, M., T. Bever and J. Fodor. 1966. The active use of grammar in speech perception. *Perception & Psychophysics* 1(1), 30-32.
- Haider, H. 2023. Is Chat-GPT a grammatically competent informant? Ms., Salzburg University.
- Howitt, K., S. Nair, A. Dods and R. M. Hopkins. 2024. Generalizations across filler-gap dependencies in neural language models. *arXiv preprint arXiv:2410.18225*.
- Hymes, D. 1972. On communicative competence. In J. B. Price and J. Holmes, eds., *Sociolinguistics: Selected Readings*, 269-293.
- Kamath, G., S. Schuster, S. Vajjala and S. Reddy. 2024. Scope ambiguities in large language models. *Transactions of the Association for Computational Linguistics* 12, 738-754.
- Kim, S.-Y. 2000. Acceptability and preference in the interpretation of anaphors. *Linguistics* 38(2), 315-353.
- Kobzeva, A., S. Arehalli, T. Linzen and D. Kush. 2025. Learning filler-gap dependencies with neural language models: Testing island sensitivity in Norwegian and English. *Journal of Memory and Language* 144, 104663.
- Mahowald, K., A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum and E. Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences* 28(6), 517-540.
- Manning, C. D. 2022. Human language understanding & reasoning. *Daedalus* 151(2), 127-138.
- Meseguer, E., M. Carreiras and C. Clifton. 2002. Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & cognition* 30(4), 551-561.
- Millière, R. 2024. Language models as models of language. *arXiv preprint arXiv:2408.07144*.
- OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Reinhart, T. 1981. Definite NP anaphora and c-command domains. *Linguistic inquiry* 12(4), 605-635.
- Shin, U., E. Yi and S. Song. 2023. Investigating a neural language model's replicability of psycholinguistic experiments: A case study of NPI licensing. *Frontiers in Psychology* 14, 937656.
- Wilcox, E. G., R. Futrell and R. Levy. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry* 55(4), 805-848.

Examples in: English and Korean

Applicable Languages: English and Korean

Applicable Level: Tertiary