Linguistic Research 42(3): 539-579
DOI: 10.17250/khisli.42.3.202512.003

Good-enough but more error-prone: Garden-path processing in GPT models[*]

Jonghyun Lee[**] · Jeong-Ah Shin[***]
(Korea University Sejong Campus · Dongguk University)

Lee, Jonghyun and Jeong-Ah Shin. 2025. Good-enough but more error-prone: Garden-path processing in GPT models. *Linguistic Research* 42(3): 539-579. This research explores the syntactic processing of Large Language Models (LLMs), specifically GPT-3.5 and GPT-4, by comparing them to human processors, focusing on garden-path sentences. These structures are challenging for even proficient human processors, often causing misinterpretations that persist despite reanalysis, revealing the 'good-enough' nature of human syntactic processing. This study aims to determine if LLMs exhibit a similar 'good-enough' syntactic processing as humans and whether more advanced models exhibit a more human-like processing. In a series of experiments, we examined how models handle garden-path sentences such as "While the man hunted the deer ran into the woods," through a comprehension questions task. A key focus was whether misinterpretations in the target phrases ("hunted the deer") erroneously affected the global interpretation of the sentence. Results showed that LLMs display patterns similar to humans, including lingering misinterpretations and the ability to utilize linguistic cues such as plausibility, phrase length, and verb type. This suggests that LLMs mimic human 'good-enough' syntactic processing through probabilistic next-word prediction, including making human-like errors. However, LLMs also showed vulnerability to garden-path structures, showing a higher rate of errors compared to humans, likely due to inherent features of their processing mechanisms. **(Korea University Sejong Campus · Dongguk University)**

Keywords ChatGPT, artificial intelligence, large language models, syntactic ambiguity, good-enough processing, garden-path sentences

[*] This work was supported by a Korea University Grant and the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2025S1A5A8005032) to Jonghyun Lee and by the Dongguk University Research Fund of 2025 to Jeong-Ah Shin.

[**] First Author

[***] Corresponding author

1. Introduction

In a short period, the advent of large language models (LLMs) has made a profound impact on the world. These models have transformed the landscape of artificial intelligence by showcasing their remarkable capacities in understanding and generating language. Contrary to their predecessors, LLMs such as GPT-4 exhibit a level of linguistic comprehension and production that often parallels, and in some cases, surpasses human abilities (Bojic et al. 2023; Herbold et al. 2023; Orru et al. 2023; Taloni et al. 2023; Wang et al. 2024; Zhai et al. 2024). This leap in performance has not only gathered widespread attention but has also reshaped our perception of AI's potential in language comprehension, challenging our previous assumptions about the limits of machine-based language processing.

While LLMs demonstrate impressive linguistic performance, the language processing mechanisms of LLMs remain largely opaque, particularly pertains to how they handle complex syntactic structures in language. This lack of clarity raises questions about the nature of their language comprehension and production—whether they process language in a manner that reflects human-like processing or if their capabilities are more akin to sophisticated algorithms with a huge amount of input data simulating language use (Linzen and Baroni 2021). This research paper seeks to explore these questions, focusing primarily on the syntactic processing of LLMs.

A deeper understanding of their mechanisms not only promises to enhance the development of LLMs by identifying and improving their linguistic limitations but also offers insights into human language acquisition and processing. This research is particularly concerned with the syntactic processing of LLMs, which, despite not being explicitly taught specific linguistic rules, often display a remarkable grasp of complex syntactic structures, mirroring human-like processing to some extent (Wilcox et al. 2018; Futrell et al. 2019; Wilcox et al. 2019; Hu et al. 2020; Linzen and Baroni 2021; Lee et al. 2022; Lee and Shin 2023). This phenomenon raises intriguing questions about the nature of language acquisition and processing, both in machines and humans. While humans rely on a rich mix of linguistic input and cognitive faculties (Hauser et al. 2002), LLMs achieve a semblance of this understanding through vast amounts of textual data and algorithmic processing. Investigating how LLMs, through sheer exposure to language data, manage to simulate an understanding of syntactic rules may offer potential insights into the learning mechanisms in both AI and humans. This research aims to shed light on these parallels and distinctions, contributing to our knowledge of language processing in both domains.

In addressing the intricate capabilities of LLMs' syntactic processing, this research employs a targeted evaluation approach, a method grounded in psycholinguistic experimental techniques (Linzen et al. 2016). This approach enables the assessment of LLMs' syntactic processing by observing their responses to carefully constructed sentences that challenge specific syntactic representations (Marvin and Linzen 2018). Similar to the methods used to unravel the complexities of human language processing, this technique allows for a nuanced examination of how LLMs navigate linguistic structures. For instance, Linzen et al. (2016) explore the ability of Long Short-Term Memory (LSTM) networks to learn syntax-sensitive dependencies, such as English subject-verb agreement. The study showed LSTMs could grasp grammar with high accuracy if trained explicitly, highlighting the need for more sophisticated architectures that benefit from direct supervision to improve the learning of syntax-sensitive dependencies. This approach may serve as a pivotal tool for probing the LLMs language processing and guiding future advancements in model architecture and training.

Following the pioneering work on the targeted evaluation approach, numerous studies have applied this methodology to explore a broad range of syntactic aspects across different LLMs, including Transformers (Bacon and Regier, 2019; Goldberg 2019; Van Schijndel et al. 2019; Ettinger 2020; Hu et al. 2020; Kuncoro et al. 2020; Lee et al. 2022; Lee and Shin 2023) as well as Recurrent Neural Networks (RNNs) (Linzen et al. 2016; Futrell et al. 2018; Gulordava et al. 2018; Marvin and Linzen 2018; van Schijndel and Linzen 2018; Wilcox et al. 2018; Frank and Hoeks 2019; Futrell et al. 2019; Wilcox et al. 2019; Chaves 2020; McCoy et al. 2020). These studies, employing diverse metrics such as accuracy (Linzen et al. 2016), surprisal (Futrell et al. 2018), and attention maps (Lee and Shin 2023), have demonstrated that LLMs can exhibit human-like syntactic processing patterns across various syntactic dimensions, including agreement (Gulordava et al. 2018), subordination (Wilcox et al. 2018), garden-path sentences (van Schijndel and Linzen 2018), and long-distance dependencies (Wilcox et al. 2018). Importantly, this syntactic proficiency is not necessarily the result of additional, syntax-focused training. While earlier models struggled with complex syntactic structures when trained solely on general linguistic tasks (Linzen et al. 2016), more advanced models trained on larger datasets showed syntactic processing often comparable to those of humans, without explicit syntactic training (Wilcox et al. 2018). This finding suggests the potential of LLMs to internalize complex syntactic knowledge from extensive data exposure alone.

However, when dissecting the syntactic abilities of individual models, it becomes evident that not all models uniformly exhibit human-like syntactic proficiency across all aspects of syntax. For instance, while BERT has shown overall proficiency in subject-verb agreement tasks, it falls short in more complex structues such as agreements within object relatives clause (van Schijndel et al. 2019). More interstingly, LLM's overall language performance does not guarantee superior syntactic processing capabilities. Even advanced models such as GPT, have not consistently surpassed LSTMs in syntactic tasks (van Schijndel et al. 2019), encountering similar challenges and showing less human-like performance in several syntactic tasks (Bacon and Regier 2019). Moreover, the significant enlargement of network size often yields only marginal syntactic performance improvements (Lee et al. 2022). This suggests that a model's general linguistic performance or size does not necessarily guarantee superior syntactic capabilities.

Yet, it remains possible that these earlier models had not reached a sufficient threshold of advancement for general improvements to translate into syntactic gains. The emergence of GPT-3.5 and GPT-4 presents an opportunity to test this possibility. Based on these models, ChatGPT has demonstrated language abilities that often approach or even surpass human performance in several language-related tasks, such as essay writing (Herbold et al. 2023) and problem solving (Orru et al. 2023; Zhai et al. 2024). These advances represent a qualitative leap beyond the Transformers examined in previous syntactic evaluations, raising the question of whether such substantial improvements in general linguistic capabilities might finally yield corresponding enhancements in syntactic processing. This study aims to explore this question, examining whether GPT-3.5 and GPT-4's exceptional general performance is paralleled by improved syntactic abilities.

Despite the growing interest in ChatGPTs, attempts to apply the targeted evaluation approach to the model for the assessment of its syntactic abilities have been scarce. A notable challenge is the lack of access to ChatGPT's code since GPT-3.5, preventing researchers from examining its internal processes with surprisal or attention maps, which is similar to analyzing online data such as reaction times or brain imaging in human studies. However, it remains possible to probe into GPTs through methods analogous to studying human language processing, such as production, accuracy assessments, or grammatical judgment tasks.

One such study by Cai et al. (2024) aims to evaluate how ChatGPT's language processing compares to that of humans, conducting 12 pre-registered psycholinguistic experiments. This suite of experiments included two specific tests designed to scrutinize its syntactic processing: syntactic priming and syntactic ambiguity resolution. These tests revealed that ChatGPT exhibits human-like patterns in syntactic reuse but it showed distinct approaches in managing syntactic ambiguities. In cases of syntactic ambiguity, humans often interpret ambiguous phrases by relying on contextual clues to resolve noun-verb (NP/VP) ambiguities, but ChatGPT did not show a preference for using contextual information in the same manner as humans. This suggests that ChatGPT's approach to resolving syntactic ambiguities may diverge from the human processing model.

Following this line of research that demonstrates the viability of production-based methods for probing LLM syntax (Bazhukov et al., 2024; Cai et al., 2024; Qiu et al., 2025), the current study applies similar approaches to examine ambiguity resolution in LLMs. Ambiguity resolution is valuable for understanding syntactic processing because it reveals the complex interplay of multiple linguistic factors during sentence comprehension. When faced with ambiguous structures, the choice between competing interpretations depends on various linguistic cues such as semantic plausibility, phrase length, verb properties, and contextual information. Previous research indicates that LLMs similarly weigh such linguistic cues, for instance considering plausibility and transitivity when resolving ambiguities, drawing parallels to human syntactic processing (Futrell et al. 2018; van Schijndel and Linzen 2018; Lee et al. 2022; Lee and Shin 2023). Thus, by examining how models resolve syntactic ambiguities, we can gain insights into which linguistic factors they prioritize and how they utilize different sources of information in syntactic processing.

Among various types of syntactic ambiguity, this study focuses specifically on garden-path sentences. Garden-path sentences are temporarily ambiguous structures that initially guide processors toward incorrect interpretations before requiring reanalysis for proper comprehension (Bever 1970; Frazier and Rayner 1982). For instance, the sentence "While the man hunted the deer ran into the wood" initially suggests "the deer" as the object of "hunted," whereas it actually serves as the subject of "ran" (Christianson et al. 2001). This setup prompts a reanalysis, where the misinterpretation of 'the deer' as the object of 'hunted' is abandoned while recognizing its true function as the subject of ‘ran.’ During this reanalysis process, ‘the deer’ is syntactically ambiguous since it can be considered either as the object of ‘hunted’ or the subject of ‘ran,’ which should be resolved to correctly parse the sentence.

This structure provides several methodological advantages over the NP/VP ambiguities examined in previous research (Cai et al. 2024). In contrast to NP/VP ambiguities where multiple interpretations remain equally valid, garden-path sentences require identification of a singular syntactically correct resolution. This characteristic enables clear assessment of accuracy, as there is a correct interpretation that can be evaluated. Furthermore, these structures present significant processing challenges for human readers as well, establishing them as a benchmark for evaluating whether LLMs can match or exceed human syntactic processing capabilities (MacDonald 2013). By choosing such tasks, we can more meaningfully discuss and evaluate the advancements in LLMs’ capabilities in handling complex syntactic structures.

In particular, this study focuses more on the second aspect, the challenging nature of processing garden-path sentences for human readers, as it could provide insightful observations relevant to both human and LLM syntactic processing, especially in the context of a “good-enough” representation. Ferreira and Patson (2007) proposed the good-enough model of language comprehension, noting that individuals frequently misinterpret garden-path sentences, as initial misinterpretations continue to affect their overall understanding. For example, upon reading a garden-path sentence such as “While Mary bathed the baby played in the crib,” the majority of readers would mistakenly respond ‘yes’ when asked if “Mary bathed the baby” (Christianson et al. 2001). These erroneous responses, however, do not necessarily indicate a failure in resolving syntactic ambiguity, because most participants also correctly answered ‘yes’ to “Did the baby play in the crib?,” revealing that they did recognize ‘the baby’ as the subject of ‘played.’ Based on these findings, Ferreira and Patson (2007) argued that humans frequently construct a quick and approximate interpretation of a sentence without engaging in deep syntactic analysis, particularly in cases of complex or ambiguous syntax. This “good-enough” approach demonstrates the fallibility of human syntactic processing—or, from another perspective, a strategic form of processing that prioritizes the efficiency (Ferreira and Patson 2007).

In this regard, this study seeks to explore how ChatGPT processes garden-path sentences. Previous research indicates that models based on the Transformer architecture exhibit processing strategies akin to those of human processors when confronted with garden-path sentences, utilizing linguistic cues such as plausibility and transitivity (Lee et al. 2022; Lee and Shin 2023). Given the advancements in GPT's general language performance beyond its predecessors, it is hypothesized that it may also exhibit a processing pattern for garden-path sentences that closely resembles human comprehension strategies. Nevertheless, it is crucial to note that superior general linguistic performance does not always imply enhanced syntactic processing (van Schijndel et al. 2019). This raises the question of whether GPT's interpretations of garden-path sentences consistently align with human processing patterns.

The complexity of predicting outcomes is further compounded when considering a human-like processing pattern of garden-path sentences does not necessarily represent the most syntactically accurate interpretation. From the perspective of the "good-enough" representation, human syntactic processing can be susceptible to errors, with syntactic ambiguity often resulting in incorrect interpretations. If LLMs indeed emulate human syntactic processing, they might also replicate these flawed syntactic processing patterns. Nonetheless, given the distinct learning mechanisms underpinning LLMs, it is plausible that they may approach the processing of garden-path sentences in a fundamentally different manner, potentially exhibiting either "mechanical" accuracy or even more probabilistic, heuristic-based processing. This research aims to explore these issues, summarizing them into the following research questions:

1. Do GPT models manifest a pattern of syntactic processing in the resolution of garden-path sentences that parallels human cognitive processes? Specifically, how do they handle syntactic constructs that typically pose challenges to human processing, and do they generate more or fewer errors in these contexts?
2. Does a GPT model with superior general performance exhibit better or more human-like syntactic processing?

To investigate these questions, this research employed a targeted evaluation approach, conducting three tests on the GPT-3.5 and GPT-4 models, which serve as the foundation for ChatGPT. Initially, the results from these models are compared with those from human parsers to perform a comparative analysis between LLMs and humans. Concurrently, a comparison between GPT-3.5 and its successor, GPT-4, which demonstrates superior general performance, is conducted to assess the impact of advancements in language model capabilities on syntactic processing efficiency.

2. Experiment 1

Experiment 1 applied the methods adapted from the first experiment of Christianson et al. (2001) to GPTs and compared these results with those obtained from human participants in Christianson et al. (2001). Christianson et al. (2001), in their first experiment, investigated participants' responses to the sentence such as "While the man hunted the deer ran into the woods" followed by the question such as "Did the man hunt the deer?". In this sentence, "the deer" could initially be misinterpreted as the object of "hunted," despite it serving as the subject for "ran." Consequently, a syntactically correct interpretation of the sentence does not provide evidence of the man hunting the deer, rendering "No" as the accurate response to the question. The study also considered two main linguistic factors within garden path sentences: length (long *vs.* short) and plausibility (plausible *vs.* implausible). It is hypothesized that longer conditions likely increase errors by extending the time readers dwell on the incorrect interpretation, while plausible conditions do so by making it harder for readers to reject the misunderstanding. The results of Christianson et al. (2001) showed that participants were significantly more likely to incorrectly answer "yes" in the garden-path condition (Figure 1). Additionally, more plausible final interpretations or longer ambiguous regions resulted in more incorrect "yes" responses. These results suggest that participants often stick with their initial misinterpretation of the sentence's meaning without fully reanalyzing its structure. Experiment 1 examined the responses of GPT-3.5 and GPT-4 to identical sentences, aiming to determine if these models exhibit the same tendency for inaccurate responses and those are influenced by plausibility and length.

2.1 Method

2.1.1 Models

For LLMs, we utilized the latest versions of GPT-3.5 and GPT-4, as available in December 2023. GPT-4 represents a more advanced model, demonstrating better contextual comprehension, greater coherence over longer passages, and enhanced accuracy in generating contextually appropriate responses (Achiam et al. 2023).

2.1.2 Procedure

The experiment was executed via OpenAI's chat completion API. During the task, the models were instructed to read the sentence and answer the subsequent question with either 'Yes' or 'No.' They were, then, presented with sentences such as (1a) and asked to respond 'yes' or 'no' to questions such as (1d). The task was repeated ten times per each sentence, and the results were averaged across the iterations. Each trial used zero-shot prompts through separate API calls where the conversation context was reset after each response, ensuring no influence between trials.

2.1.3 Materials

The study employed the same 42 sentence items used in Christianson et al. (2001). Each item was presented under one of six conditions as in (1). While typically human participants would encounter only one version of each sentence item, in this task, the models were exposed to all six conditions. However, as noted in the procedure section, the experiment was conducted through separate API calls, which ensured that responses to previous items did not influence subsequent ones. The garden-path sentences were manipulated across four conditions by altering two linguistic factors: the *length* of the ambiguous region (short, "the deer," *vs.* long, "the deer that was brown and graceful") and the *plausibility* of misinterpreting the ambiguous noun phrase (NP) as the object of the main verb (plausible, "the hunted deer ran into the woods," *vs.* implausible, "the hunted deer paced in the zoo"). Non-garden path sentences were differentiated from garden path sentences through the inclusion of an additional NP (e.g., "the pheasant"), with two non-garden path conditions varying only in the length of the ambiguous region. The same question was presented for all six versions of a sentence item such as (1d).

(1) a. While the man hunted the deer (that was brown and graceful) ran into the woods. [Garden-path − Plausible]
b. While the man hunted the deer (that was brown and graceful) paced in the zoo. [Garden-path − Implausible]
c. While the man hunted the pheasant the deer (that was brown and graceful) ran into the woods. [No Garden-path − Plausible]

d. Question: Did the man hunt the deer?

* Inclusion of the word in parentheses indicates long conditions.

2.1.4 Statistical analysis

The statistical analysis was conducted using the R programming environment (R Core Team 2023), utilizing generalized linear mixed-effects models (GLMM) (Baayen et al. 2008) from lmerTest (Kuznetsova et al. 2017) packages to evaluate differences in comprehension question accuracy. Initially, we aimed to identify difference in the error rates between the garden-path and non-garden-path conditions by incorporating Garden-path and Model as fixed effects and Items as random effects, leading to a 2×2 factorial analysis (Garden-path *vs.* Non-garden-path × GPT-3.5 *vs.* GPT-4). The second phase focused on the garden-path sentences, analyzing the effects of length and plausibility. The model assumed Plausibility, Length, and Model as fixed effects and Items as random effects, forming 2×2×2 factorial design (Plausible *vs.* Implausible × Short *vs.* Long × GPT-3.5 *vs.* GPT-4). The third phase assessed error rate differences in non-garden-path sentences, with Length and Model as fixed effects, and Items as random effects, in a 2×2 factorial design (Short *vs.* Long × GPT-3.5 *vs.* GPT-4). In all analyses, when any interaction was detected, post-hoc analyses were conducted, using the emmeans R package (Lenth 2023). In instances of convergence issues or failure to meet the assumption of multicollinearity, the strategy was to start by removing the random intercept or the three-way interaction, potentially simplifying the model further by eliminating other interactions if necessary, and then proceeding with the statistical analysis using these reduced models. Any employment of a simpler model was stated in the results. Across all analyses, the dependent variable was coded as a binary outcome, with 0 and 1, where 1 denoted an erroneous response, specifically "Yes" in this experiment. All human data were obtained from Christianson et al. (2001). Given that the human data consist of collective responses from numerous individuals, while the LM data results from multiple iterations by a singular model, this discrepancy may affect the suitability of inferential statistics for head-to-head analysis. Consequently, our analysis between humans and LMs has been oriented towards utilizing descriptive statistics to highlight observable patterns, focusing on distinctions that are numerically significant. For data visualization, the Python packages seaborn (Waskom 2021) and matplotlib (Hunter 2007) were employed.

2.2 Results

In the first analysis for error rates between garden-path and non-garden-path sentences, we found significant main effects for both Garden-path (estimate=2.20, SE=0.10, z=22.75, p<0.001) and Model (estimate=1.11, SE=0.09, z=11.82, p<0.001). The results indicates that both models made more errors on garden-path sentences and GPT-3.5 had higher overall error rates than GPT-4. Additionally, a significant Garden-path × Model interaction emerged (estimate=-1.10, SE=0.09, z=-11.78, p<0.001). This interaction reveals that the two models differed most on non-garden-path sentences, where GPT-3.5 showed higher error rates than GPT-4, while their performance converged on garden-path sentences.

The second analysis, which focused on error rates within garden-path sentences, revealed significant main effects of Plausibility (*estimate*=-1.14, *SE*=0.09, z=-12.94, p<0.001), Length (*estimate*=0.85, *SE*=0.09, z=9.73, p<0.001), and Model (*estimate*=-0.25, *SE*=0.09, z=-2.92, p<0.01). These effects indicate that errors were more frequent in implausible sentences, sentences with longer ambiguous regions, and for GPT-3.5. However, given that the actual mean error rates were quite similar between GPT-4 (79.76%) and GPT-3.5 (79.88%), the main effect of Model seems to be better understood through its interactions with other factors. The analysis also revealed significant two-way interactions: Plausibility × Model (*estimate*=0.33, *SE*=0.09, z=3.81, p<0.001), where GPT-4 shows a larger plausibility effect than GPT-3.5; Length × Model (*estimate*=-0.39, *SE*=0.09, z=-4.47, p<0.001), where GPT-4 shows a larger length effect than GPT-3.5; and Plausibility × Length (*estimate*=-0.25, *SE*=0.09, z=-2.89, p<0.01), where the length effect is stronger for plausible sentences than implausible sentences. Moreover, a significant three-way interaction among Plausibility, Length, and Model was observed (*estimate*=0.17, *SE*=0.09, z=2.00, p=0.046), where GPT-4 produced more errors on long plausible sentences but fewer errors on short implausible sentences.

In the third phase of analysis for non-garden-path conditions, a significant main effect was observed for Model (*estimate*=4.20, *SE*=0.52, z=8.03, p<0.001), with GPT-3.5 showing significantly higher error rates than GPT-4. Additionally, a significant interaction between Length × Model emerged (*estimate*=0.54, *SE*=0.22, z=2.50, p<0.05), where GPT-3.5 showed higher error rates for longer sentences while GPT-4's error rates were unaffected by sentence length.

Summarizing, both models demonstrated a higher propensity for errors in garden-path conditions, influenced by sentence plausibility and length, with errors increasing in plausible and longer conditions. While the overall patterns were similar between the models, GPT-3.5 had markedly higher error rates in non-garden-path conditions. Another difference was that GPT-4 displayed a larger plausibility effect and a greater sensitivity to sentence length within garden-path conditions, particularly showing higher error rates in long sentence conditions compared to GPT-3.5.

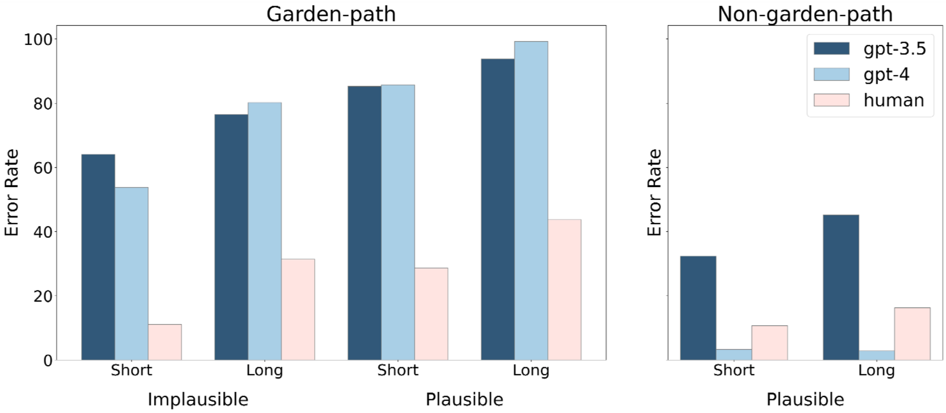

Figure 1. Error rates by Garden-path, Plausibility and Length from Experiment 1 comparing GPT-3.5, GPT-4, and humans when responding to comprehension questions, with human data referenced from Experiment 1B in Christianson et al. (2001)

When these results are compared to human performance, a similar trend emerges. Both humans and LLMs showed increased errors in garden-path conditions, influenced by plausibility and length. However, the error rates for LLMs were notably higher, especially in garden-path conditions where their error rates were more than double those observed in humans. Regarding non-garden-path conditions, the performance varied between the models; GPT-3.5 recorded much higher error rates than humans, whereas GPT-4 demonstrated comparable performance with few errors.

2.3. Discussion

Experiment 1 investigated the processing of garden-path sentences by GPT-3.5 and GPT-4, comparing their responses with human patterns. Essentially, both models showed similar sentence processing patterns to humans, more frequently producing errors in garden-path than non-garden-path sentences. These errors increased in the conditions where the initial misinterpretations were more difficult to be rejected due to increased plausibility or extended ambiguous regions. This pattern is consistent with the good-enough processing framework (Ferreira and Patson, 2007), where initial misinterpretations persist and influence sentence comprehension despite structural reanalysis. However, LLMs produced a notably higher count of erroneous responses than humans, especially in garden-path conditions. This suggests that LLMs, exhibiting syntactic processing patterns akin to humans, similarly struggle with sentences that present syntactic challenges to humans, often to an even greater extent.

In comparing GPT-3.5 and GPT-4, differences in handling garden-path sentences were minimal. GPT-4, however, did show fewer errors in implausible short conditions compared to GPT-3.5, although this advantage did not extend broadly across all conditions. Intriguingly, GPT-4 was more susceptible to incorrect "Yes" responses in scenarios involving longer ambiguous regions. GPT-4's standout performance was evident only in non-garden-path sentences, where its responses were almost error-free, surpassing even that of human participants. This contrasts with GPT-3.5, which, in non-garden-path conditions, recorded higher error rates than humans.

While GPT-3.5 showed near chance-level performance in non-garden-path conditions, the systematic variation across conditions and the low error rates observed in Experiments 2 and 3 suggest that the errors may not be attributed simply to a failure to comprehend the sentences. Previous research has reported a yes-bias in humans (Christianson et al., 2001). Given that LLMs tend to show a preference for affirmative responses to input, it is possible that a similar yes-bias exists in these models. However, as will be shown in Experiment 3, the low error rates observed in certain conditions indicate that the errors cannot be attributed solely to yes-bias.

3. Experiment 2

Experiment 2 adapted methods from Christianson et al. (2001)'s second experiment, which aimed to explore three issues raised from their first experiment.

The first issue Christianson et al. (2001) identified was whether the "yes" responses in their experiment resulted from a lingering misinterpretation despite syntactic reanalysis, or from a mere failure to reanalyze. For instance, in the sentence “While the man hunted the deer ran into the wood,” there is a possibility that “the deer” incorrectly parsed as the object of “hunted” and the subject of “ran” was erroneously left empty. Their second experiment probed this by asking questions such as “Did the deer run into the woods?” and seeing if participants correctly identified subjects of the main clause, which indicates a reanalysis from their initial misunderstanding.

The second issue examined whether the errors in garden path sentences stemmed from an initial syntactic error or from pragmatic expectations. The participants may interpret “the man” as hunting “the deer”, not because they were led down the “garden path” to an incorrect initial interpretation but because the real-world likelihood suggest that “the man” would be hunting “the deer.” By reversing the order of clauses in sentences (e.g., “The deer ran into the woods while the man hunted”), the experiment tested if participants would still assume “the man” was hunting “the deer” without the syntactic cues leading to this initial garden path misinterpretation, thus differentiating between syntactic influence and pragmatic inference.

The third issue explored how the length of an ambiguous sentence region affected interpretation, questioning if the challenge lied in the duration of holding onto an incorrect role assignment or in the overall sentence length. The latter hypothesis gained some support from observations that even a non-garden path condition showed length effects, suggesting that sentence length itself could contribute to interpretation difficulty. The experiment further investigated this by manipulating the position of the head of the ambiguous noun phrase (“the deer that was brown and graceful” to “the brown and graceful deer”), offering another way to examine the impact of committing to a wrong analysis over time.

Our experiment 2 investigated how GPT-3.5 and GPT-4 perform the same three issues identified by Christianson et al. (2001). In Experiment 1, the models exhibited error patterns similar to humans, and under garden-path conditions, LLMs were more likely to incorrectly respond with "yes," suggesting a potential reanalysis failure. Experiment 2 first sought to determine whether the models could correctly capture the target word as the subject in the main clause, using targeted questions such as (2f). Additionally, Experiment 2 aimed to discern whether the high error rates were due to purely syntactic misinterpretations or if pragmatic factors also influenced the outcomes. Given that LLMs learn language from extensive text data, not syntactic rules, their sentence processing might lean more on world knowledge. This is partially supported by GPT-3.5's high error rates in non-garden-path sentences, which may indicate a reliance on pragmatic interpretation. The experiment also investigated into the underlying causes of length effect. The models showed more errors in longer sentences, suggesting a heightened sensitivity to sentence length. This is intriguing since transformer-based LLMs do not process sentences word-by-word manner, but through an attention mechanism that assesses the entire input at once. Understanding why GPTs are impacted by sentence length could reveal deeper insights into their limitations and processing strategies.

3.1 Method

3.1.1 Models and procedure

The models and procedure were identical to those in Experiment 1. Each sentence was tested 10 times, with results averaged across iterations

3.1.2 Materials

Experiment 2 utilized 40 sentence items identical to those used in Christianson et al. (2001)'s Experiment 2. Following Christianson et al. (2001)'s design, these sentences were presented across different experimental conditions with modifications to address three issues. The first adjustment involved introducing a new condition where the question targeted the interpretation of the main clause, exemplified by (2f) (Main Clause Probing Question). Furthermore, the order of clauses was altered, arranging the subordinate clause either before (Garden-path) or after (Non-garden-path) the main clause. Lastly, the positioning of the head of the ambiguous noun phrase was varied, occurring either early (Head Early) or late (Head Late) within the phrase.

(2) a. While the man hunted the deer that was brown and graceful ran into
the woods. [Garden-path – Head Early]
b. While the man hunted the brown and graceful deer ran into the woods.
[Garden-path – Head Late]
c. The deer that was brown and graceful ran into the woods while
the man hunted. [Non-garden-path – Head Early]

d. The brown and graceful deer ran into the woods while the man hunted. [Non-garden-path – Head Late]

e. Subordinate Clause Probing Question: Did the man hunt the deer?

f. Main Clause Probing Question: Did the deer run into the woods?

In the human study by Christianson et al. (2001), sentence items and question types were counterbalanced using a Latin square design, such that each participant saw each sentence item only once but encountered both subordinate and main clause probing questions across different items. In contrast, the current LLM experiment tested each model on all conditions for every sentence item, as the separate API calls ensured that previous trials did not influence subsequent responses.

3.1.3 Statistical analysis

All statistical analysis procedures were the same as those in Experiment 1 and were conducted in two stages. The initial phase aimed to discern differences attributable to question type and the influence of LLMs. In this stage, Question Type and Model were treated as fixed effects, with Items as random effects, forming a 2×2 factorial design (Subordinate Question vs. Main Question × GPT-3.5 vs. GPT-4). The subsequent phase focused on the subordinate clause probing questions, structured in a 2×2×2 factorial design (Garden-path vs. Non-garden-path × Early vs. Late × GPT-3.5 vs. GPT-4) with Garden-path, Head Position, and Model as fixed effects and Items as random effects. In each of the analyses, the dependent variable was encoded as a binary outcome, represented by 0 and 1. Here, the value 1 indicated an incorrect response: "Yes" in the context of subordinate clause probing questions and "No" for main clause probing questions.

3.2 Results

In the first analysis examining variance in error rates between subordinate and main clause probing questions, we identified significant main effects for Question Type (estimate=-4.36, SE=0.18, z=-24.38, p<0.001) and Models (estimate=0.99, SE=0.15, z=6.45, p<0.001), along with a significant interaction between Question Type and

Models (estimate=0.67, SE=0.15, z=4.36, p<0.001). The analysis revealed that both models were more prone to errors with main clause questions compared to subordinate clause questions, with GPT-3.5 generating significantly more errors than GPT-4. The interaction indicated that GPT-3.5's higher error tendency was more pronounced for main clause questions than subordinate clause questions.

The second analysis focusing on the subordinate questions revealed significant main effects of Garden-path (estimate=0.97, SE=0.11, z=9.19, p<0.001), Head Position (estimate=0.79, SE=0.10, z=7.61, p<0.001), but no significant main effect of Model (p<0.1). Garden-path sentences and early head sentences produced significantly more errors than their respective counterparts. Additionally, three significant two-way interactions were observed. The Garden-path × Head Position interaction (estimate=0.54, SE=0.10, z=5.21, p<0.001) indicated that the head position effect was stronger for garden-path than non-garden-path sentences. The Garden-path × Model interaction (estimate=-1.04, SE=0.11, z=-9.85, p<0.001) revealed that the garden-path effect differed between models: GPT-3.5 showed no significant difference between garden-path and non-garden-path conditions, while GPT-4 showed a strong garden-path effect with significantly higher error rates in garden-path conditions. The Head Position × Model interaction (estimate=-0.26, SE=0.10, z=-2.57, p<0.05) showed that model differences were larger for late head sentences than early head sentences. Lastly, there was a significant three-way interaction (estimate=-0.20, SE=0.10, z=-2.00, p=0.046), which indicated that the Garden-path × Head Position interaction differed between models. Specifically, GPT-4 showed a stronger head position effect in garden-path sentences compared to GPT-3.5, while both models showed similar weak head position effects in non-garden-path sentences.

In summary, both models exhibited significantly lower error rates in main clause probing questions, with the rates nearing zero (GPT-3.5: 4.06%, GPT-4: 0.19%). Between the two models, GPT-4 demonstrated significantly lower error rates compared to GPT-3.5. Additionally, an effect of Garden-path was noted in subordinate clause questions, where both models made more errors when the subordinate clause came before the main clause rather than after. However, this pattern was more evident in GPT-4, with GPT-3.5 showing no significant difference between garden-path and non-garden-path conditions. Finally, an early head position led to increased error rates in both models, with this effect being stronger in garden-path conditions overall, and this enhancement being particularly pronounced for GPT-4 compared to GPT-3.5.

When compared to human data, LLMs showed several similar patterns. First, both humans (Christianson et al. 2001) and models committed fewer errors in answering main clause questions. For subordinate clause questions, the incorrect answers significantly increased in the garden-path condition than in the non-garden-path condition. Additionally, errors were more frequent when the syntactic head appeared early in the sentence rather than later, with this effect larger for garden-path conditions. Nonetheless, a distinct pattern was observed when analyzing GPT-3.5, which differed from the human data. For this model, the error rates for subordinate clause questions were similarly high both for garden-path and non-garden-path conditions. Furthermore, the error rates for both models were notably higher than humans, as similarly observed in Experiment 1, across all four subordinate clause question conditions. On the other hand, regarding main clause questions, the models, especially GPT-4, which displayed an error rate nearly at zero, tended to have lower error rates than those recorded for humans.

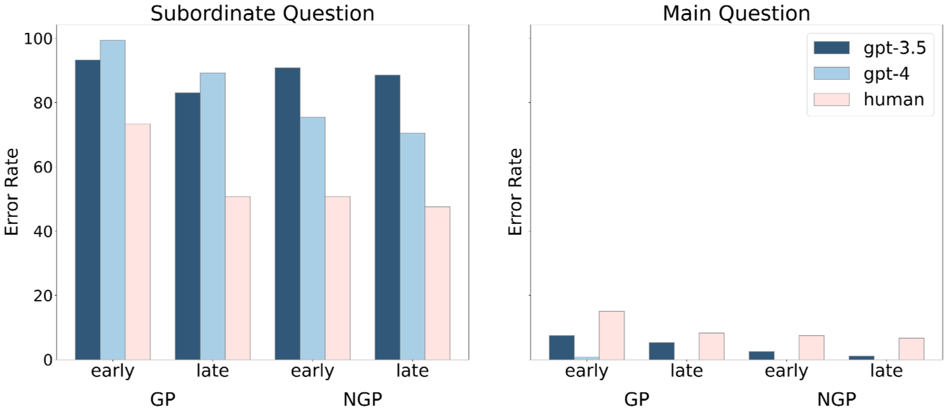

Figure 2. Error rates by Question Type, Garden-path and Head Position from Experiment 2 comparing GPT–3.5, GPT–4, and humans when responding to subordinate and main questions, with human data referenced from Experiment 2 in Christianson et al. (2001). GP = Garden-Path, NGP = Non-Garden-Path

3.3 Discussion

Experiment 2 addressed three issues identified in Experiment 1: the potential for syntactic reanalysis failure, the influence of pragmatic expectations, and the impact of head position. The findings indicated that similar to Experiment 1, LLMs generally exhibit syntactic processing that mirrors human patterns but displayed a considerably higher overall error rate than human processors.

First, the results showed that the tendency to incorrectly answer “yes” in Experiment 1 might not stem solely from the failure of reanalysis, as LLMs recognized “the deer” as the subject of the main clause, evidenced by significantly lower errors in main clause probing questions. Moreover, errors in the main clause were almost near zero, surpassing human performance. This finding aligns with good-enough processing, where initial misanalyses persist despite successful syntactic reanalysis (Christianson et al. 2001).

Secondly, the experiment demonstrated that the tendency to incorrectly answer “yes” in Experiment 1 might not arise only from pragmatic inference but also from initial syntactic difficulties, as shown by a main effect of Garden-path within subordinate clause questions. However, the observation of a numerically high error rate in non-garden-path sentences (GPT-3.5: 89.63%, GPT-4: 73.00%) indicated that pragmatic expectations also played a substantial role in the erroneous responses. Human data showed a similar trend, with significantly more errors in garden-path conditions, while also revealing error rates close to 50% in non-garden-path conditions. This suggests that their errors were likely due to a combination of syntactic misanalysis and the pragmatic plausibility of the inference.

Third, the results indicated that the length effect observed in Experiment 1 was not merely due to increased phrase length but also related to syntactic structure, as demonstrated by the head position effect, aligning with the patterns in human data. However, it remains uncertain whether this effect holds the same implications for LLMs as it does for humans. Assuming humans mostly process a sentence sequentially from its beginning, “the deer” in an early head position is likely to be assigned as the object of the verb in the subordinate clause longer than in a late position, possibly triggering the head position effect. In contrast, since LLMs process all sentence elements concurrently, an early head position does not necessarily entail prolonged assignment as the object. Instead, in LLMs, the distance between the verb and the target noun might have influenced the results, given the attention mechanism in Transformers, which tends to focus on closer words first (Clark et al. 2019), potentially causing increased errors in early head positions.

In the comparison of the models in Experiment 2, two key distinctions between

GPT-3.5 and GPT-4 were highlighted. First, in the main clause probing questions, GPT-4 demonstrated superior performance with almost no errors, compared to GPT-3.5. Yet, it is important to note that GPT-3.5 also produced minimal errors in the main clause questions, indicating that its errors were not simply due to a failure in syntactic reanalysis. Secondly, a notable difference between the two models was observed in the issue related to pragmatic expectation. GPT-4 displayed a pattern similar to humans, indicating that the incorrect answers in Experiment 1 were due to a combination of syntactic misanalysis and the pragmatic plausibility of the inference. In contrast, GPT-3.5 showed no difference between garden-path and non-garden-path conditions, differing from human patterns, which might suggest that a significant portion of its errors could be due to pragmatic inference rather than syntactic misinterpretation. Overall, in Experiment 2, GPT-4 demonstrated more human-like processing and achieved lower overall error rates compared to GPT-3.5 in contrast to Experiment 1, where the performance of GPT-3.5 and GPT-4 was nearly identical. However, it is noteworthy that GPT-4 still exhibited higher error rates in the garden-path conditions of subordinate questions.

4. Experiment 3

Experiment 3, based on Christianson et al. (2001)'s third experiment, employed two modifications to the sentence structures used in Experiment 2, specifically designed to minimize errors that could arise from pragmatic reasoning.

The first modification involved the introduction of Reflexive Absolute Transitive (RAT) verbs into the sentence structures. RAT verbs, typically associated with personal hygiene activities such *wash, bathe,* and *shave*, are grammatically structured to be understood reflexively, in the absence of a direct object. This characteristic distinguishes them from Optionally Transitive (OT) verbs such as *hunt,* used in previous experiments. For instance, 'Mary bathed' is automatically interpreted as 'Mary bathed herself,' contrasting with 'Mary hunted,' which does not imply 'Mary hunted herself.' This distinction is crucial for the experiment, as it reduces the likelihood of misinterpreting the verb's object as an unspecified, general object, thereby minimizing pragmatic-driven errors.

The second modification was the inclusion of a disambiguating comma after the verb in the subordinate clause. This comma serves as a syntactic signal to avoid garden-path misinterpretations, allowing processors to parse the sentence structure more accurately from the outset. In contrast to the clause order manipulation used in Experiment 2, which could potentially influence focus or memory retention, the comma insertion is a minimal intervention that maintains the sentence's original word order and integrity.

In Experiment 2, GPT-3.5 and GPT-4 both showed a high rate of incorrect responses in non-garden-path sentences, though these were fewer than in garden-path sentences, suggesting that pragmatic reasoning contributed considerably to the errors. Furthermore, the lack of a significant difference between garden-path and non-garden-path sentences in GPT-3.5 might indicate even stronger influences of pragmatic expectations on it. Experiment 3 is designed to determine the extent of errors attributable to syntactic misinterpretations after reducing such general reasoning with the two specific modifications.

These two modifications also serve as means to explore different aspects of the syntactic abilities of LLMs. RAT verbs, which are inherently understood reflexively without a direct object, provide a test case. If LLMs do not recognize this reflexive usage, we might observe error rates similar to those with OT verbs in the earlier experiments. Conversely, a proper understanding and application of this feature by LLMs should lead to a reduction in errors for RAT verbs. Furthermore, these changes allow for the exploration of how minimal interventions such as the addition of a comma can affect the syntactic processing of LLMs. Although LLMs do not process sentences in a strictly linear fashion, information about the position of words and phrases in a sentence is still crucial for their processing in attention mechanisms (Vaswani et al. 2017). Experiment 3 examined how such changes in structure can guide the models' interpretation of sentence structures.

4.1 Method

4.1.1 Models and procedure

The models were the same as those in Experiment 1. The procedure was also identical to that of Experiment 1, with the only difference being that each sentence was iterated

30 times due to the smaller number of materials.

4.1.2 Materials

Experiment 3 utilized 24 sentences, sourced from Christianson et al. (2001). 12 with an OT verbs such as (3) and 12 with an RAT verb such as (4). Each sentence item appeared in one of 4 conditions: garden-path, non-garden-path with comma and non-garden-path with order shift.

(3) a. While the man hunted the deer that was brown and graceful ran into the woods. [OT verbs - Garden-path]
b. While the man hunted, the deer that was brown and graceful ran into the woods. [OT verbs − Comma]
c. The deer that was brown and graceful ran into the woods while the man hunted. [OT verbs − Order]
d. Question: Did the man hunt the deer?
(4) a. While Jim bathed the child that was blond and pudgy giggled with delight. [RAT verbs - Garden-path]
b. While Jim bathed, the child that was blond and pudgy giggled with delight. [RAT verbs − Comma]
c. The child that was blond and pudgy giggled with delight while Jim bathed. [RAT verbs − Order]
d. Question: Did Jim bathe the child?

4.1.3 Statistical analysis

All statistical analysis procedures were identical to those in Experiment 1 and were conducted in two stages. In the first phase, a comparison was made between the two non-garden-path conditions to understand how comma disambiguation and clause order disambiguation impact the process of disambiguation. The analysis was structured in a 2×2×2 factorial design (RAT *vs.* OT × Comma *vs.* Order × GPT-3.5 *vs.* GPT-4), with Verb Type, Disambiguation, and Model as fixed effects and Items as random effects.

The second phase aimed to explore how the type of verb influences the processing of garden-path sentences and how this effect varies across different LLMs. This analysis was limited to garden-path sentences and non-garden-path sentences that employed comma disambiguation. This model considered Verb Type, Garden-path, and Model as fixed effects, with Items as random intercepts, structured in a 2×2×2 factorial design (RAT *vs.* OT × Garden-path *vs.* Comma × GPT-3.5 *vs.* GPT-4).

4.2 Results

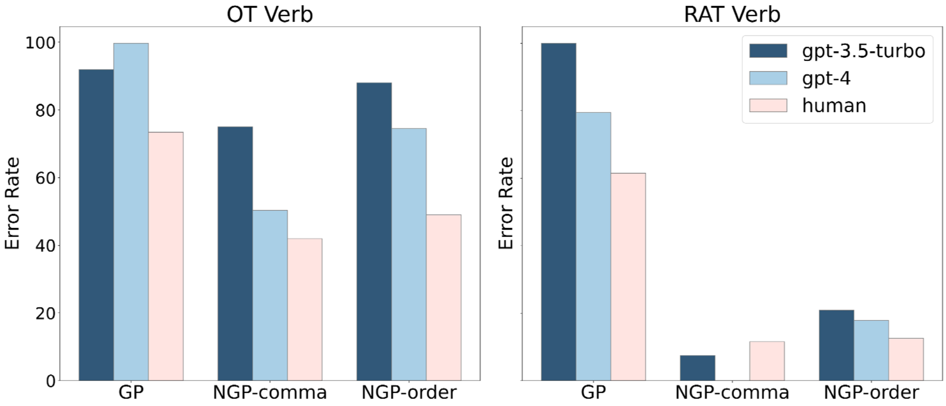

Figure 3. Error rates by Verb Type, Garden-path and Disambiguation from Experiment 3 comparing GPT-3.5, GPT-4, and humans. The human data is sourced from Experiments 3a and 3b in Christianson et al. (2001), with the error rate for garden-path conditions being an average from both 3a and 3b. GP = Garden-Path, NGP = Non-Garden-Path

In the initial analysis phase, we focused on the impact of the comma on interpreting non-garden-path sentences. Due to issues with convergence and multicollinearity, the model was simplified by removing the three-way interaction. We observed main effects for Ambiguity (estimate=-2.12, SE=0.27, z=-7.96, p<0.001), Verb Type (estimate=7.43, SE=2.04, z=3.64, p<0.001), and Model (estimate=1.79, SE=0.26, z=6.79, p<0.001), indicating more errors with comma disambiguation, OT verbs, and GPT-3.5 compared to unambiguous conditions, RAT verbs, and GPT-4, respectively. There were also significant interactions between Model and Verb Type (estimate=0.36, SE=0.09, z=3.85, p<0.001) and between Model and Ambiguity (estimate=1.27, SE=0.26, z=4.81, p<0.001). These interactions suggest that the difference between models is greater for OT verbs than RAT verbs, and greater in comma conditions than order conditions.

In the second analysis, the focus was on how the type of verb influences the processing of garden-path sentences across different LLMs. A main effect of Garden-path (estimate=13.05, SE=3.93, z=3.33, p<0.001) and a significant interaction between Verb Type and Garden-path (estimate=-9.66, SE=3.92, z=-2.46, p<0.05) were identified. These effects indicated more errors in garden-path conditions and that verb type differences emerged only in comma conditions, where OT verbs produced more errors than RAT verbs. Two marginal effects were also found. GPT-3.5 produced more errors than GPT-4 (estimate=7.34, p=0.056), and a marginal Verb Type × Model interaction (estimate=-7.12, p=0.063) showed that only GPT-4 demonstrated fewer errors with RAT verbs.

To summarize, both models exhibited a lower error rate with comma disambiguation than with order disambiguation. Furthermore, in non-garden-path conditions, both models reduced their error rates with RAT verbs as opposed to OT verbs, yet there was no observable difference in error rates between the two verb types in garden-path conditions. While the overall performance patterns were similar for both models, GPT-4 outperformed GPT-3.5 in non-garden-path conditions. Additionally, when comparing the two types of disambiguation, a larger difference was noted in the models' error rate with comma than clause order disambiguation.

Comparing these findings with human data indicates a broadly similar trend. In Christianson et al. (2001), humans also demonstrated fewer errors with RAT verbs compared to OT verbs, particularly in non-garden-path scenarios, regardless of the disambiguation method used. This pattern implies that humans may be less dependent on general reasoning, instead following structural guidelines when interpreting RAT verbs. However, humans still encountered misinterpretations with RAT verbs in garden-path conditions, suggesting a tendency toward persistent misinterpretation. Still, both models exhibited more errors than humans in garden-path conditions. The effect of commas appeared to be relatively more pronounced in LLMs. GPT-4 showed about a 15% difference in error rates between comma and clause order disambiguation, which was relatively larger than what was observed in humans. Additionally, GPT-3.5 exhibited a higher number of errors in clause order disambiguation for RAT verbs compared to humans, yet it showed fewer errors in comma disambiguation. However, since Christianson et al. (2001) did not conduct a statistical comparison of the two disambiguation methods, so this finding should be considered an observed trend.

4.3 Discussion

Experiment 3 investigated whether LLMs continue to produce incorrect responses in garden-path processing even when sentence modifications are designed to minimize errors stemming from pragmatic reasoning.

The findings indicated that despite modifications to reduce pragmatic reasoning influences, both LLMs still committed errors in comprehending garden-path sentences, mirroring human participants' behavior. The models exhibited fewer errors with RAT verbs compared to OT verbs in non-garden-path conditions, suggesting they could adequately interpret and apply the syntactic properties of RAT verbs. However, they still encountered significant errors in garden-path sentences with RAT verbs, comparable to those with OT verbs, indicating that errors were largely due to lingering syntactic misinterpretations.

While the general trend in processing garden-path sentences was aligned with human data, LLMs consistently recorded higher error rates than humans. GPT-4 demonstrated a decrease in errors when handling RAT verb garden-path sentences, yet its error rate remained higher compared to humans, who similarly exhibited lower error rates in these scenarios.

In the comparison between the two models, both showed comparable patterns, yet GPT-4 aligned more closely with human-like processing and incurred fewer mistakes overall. Moreover, GPT-4 demonstrated a lower error rate in non-garden-path conditions, regardless of the disambiguation types. Hence, while GPT-3.5 utilizes linguistic cues such as RAT verbs and commas similarly to GPT-4, it seems to leverage these cues less effectively than GPT-4.

5. General discussion

This study explored the intricacies of how GPT models process garden-path sentences, drawing a comparison with human syntactic processing to enhance our understanding of both machine and human language processing. Our investigation centered on two key questions: whether GPT models manifest a human-like 'good-enough' processing in resolving garden-path sentences, and if a model with superior general linguistic performance demonstrates enhanced syntactic processing as well. The answer to the first question is "yes" — LLMs demonstrated 'good-enough' patterns akin to humans, yielding even higher error rates. Regarding the second question, the answer is only partly "yes" as GPT-4 showed overall more human-like performance compared to GPT-3.5, particularly in unambiguous conditions, but failed to demonstrate such superiority in garden-path conditions. We will explore this in greater detail below, organizing the analysis into four distinct sections.

5.1 Similar good-enough processing and the role of probabilistic mechanisms

LLMs demonstrated patterns in processing garden-path sentences remarkably similar to humans, producing similar errors as well. LLMs more frequently answered "yes" for garden-path compared to non-garden-path sentences, but these errors were not due to a failure in reanalyzing ambiguous noun phrases as the subject of the main clause and not merely induced by pragmatic inference. Furthermore, human-like patterns in error responses influenced by length of ambiguous regions, plausibility, head position, and verb type were also observed in LLMs. They made more errors in sentences with longer ambiguous regions, higher plausibility, early head positions, and with OT verbs. These findings suggest that LLMs showed human-like 'good-enough' processing patterns in handling garden-path sentences.

The observation that LLMs exhibit patterns akin to those of humans, including errors, raises questions about the processing nature of both humans and LLMs. The similarity in outcomes does not necessarily equate to identical internal structures or mechanisms behind these patterns. Particularly, from the perspective that differentiates between performance and competence, similar performance levels do not automatically indicate similar cognitive or knowledge systems. However, the ability of LLMs to mirror certain aspects of human syntactic behavior suggests that at least some facets of human syntactic processing might be computationally replicable, without relying on the concept of unique syntactic capabilities inherent to humans.

The core mechanism underlying this computational replicability is probabilistic learning through next-word prediction. This essential mechanism has remained consistent since the older Simple Recurrent Network (SRN) adapted by Elman (1990), which was a pioneering way for neural networks to handle sequential data. It was designed to predict the next word based on the current word and the network's memory of previous words. Transformers, which GPTs are based on (Radford et al. 2019), also learn by predicting the next item in a sequence and adjusting their internal parameters to minimize the error in next-word prediction (Vaswani et al. 2017). Certainly, Transformers employ a more advanced approach, or attention mechanism, which allows the model to weigh different parts of the input sequence differently, taking into account the entire context and not just the recent past. Yet, fundamentally, their mechanisms hinge on calculating probabilities through next-word prediction and updating these probabilities accordingly (Suresh et al. 2023).

This probabilistic learning mechanism might naturally give rise to good-enough processing patterns. When processing garden-path sentences through next-word prediction, LLMs form locally coherent structures based on high-probability word sequences in their training data. For instance, in "While the man hunted the deer ran into the woods," the sequence "hunted the deer" represents a highly probable verb-object relationship. Once formed, these probabilistically-weighted initial interpretations persist even when later context (e.g., "ran") signals the need for reanalysis, because the model has already committed substantial computational weight to the initial high-probability parse. This might create the characteristic good-enough pattern observed in our experiments. Thus, good-enough processing in LLMs emerges not as a processing failure but as an inherent consequence of learning and processing language through probabilistic constraints.

If probabilistic constraints can account for good-enough processing in LLMs, similar mechanisms might underlie human syntactic processing as well. This notion of explaining human language acquisition and processing through probabilistic constraints or statistical learning is not novel. Bever (1970), who spurred research into the processing of Garden-path sentences with the famous "The horse raced past the barn fell," had already introduced the concept of probabilistic constraints and an emergentist approach to language development. In the psycholinguistic tradition stemming from Bever's ideas, the statistical patterning of words in language (and other nonlinguistic inputs to the cognitive system) is foundational to linguistic competence and, indeed, performance. In this tradition, the "meaning" of a word is distilled to a statistical analysis of the contexts in which the word appears (Altmann 2013). The findings of this study could serve as supportive evidence for this tradition, demonstrating that probabilistic learning mechanisms can computationally replicate human-like 'good-enough' syntactic processing patterns.

However, while probabilistic learning can explain good-enough processing in LLMs and potentially in humans, equating their mechanisms requires careful consideration of the vast disparity in input data between LLMs and humans. The evolution of language models into LLMs is more than the sophistication of algorithms such as the attention mechanism; often considered more crucial is the sheer volume of data these models learn from (Kaplan et al. 2020). GPTs utilize the decoder part of the Transformers architecture (Vaswani et al. 2017) and there has not been a revolutionary change in this structure since GPT-2 (Radford et al. 2019; Brown et al. 2020; Achiam et al. 2023). In contrast, the amount of data and the size of models have exponentially increased, with GPT-3 known to be trained on billions of data points and parameters (Brown et al. 2020), and GPT-4 likely using even more. On the other hand, even if humans are assumed to be engaged in language activities around the clock, the total exposure to language for one individual in a year would be around 100 million words, which is a fraction of what LLMs are exposed to.

This disparity suggests that even if humans do rely on probabilistic mechanisms similar to LLMs, they acquire and utilize probabilistic constraints with remarkable efficiency. Humans, particularly infants, demonstrate sensitivity to statistical patterns with minimal exposure. For instance, 8-month-old infants can extract statistical regularities from continuous speech after only brief exposure (Saffran et al. 1996), and young children can make sophisticated inductive generalizations by considering both samples and sampling processes (Gweon et al. 2010). This efficiency stands in contrast to the massive data requirements of LLMs to achieve comparable good-enough processing patterns. The question, then, is not simply whether humans use probabilistic learning, but rather what mechanisms enable humans to learn and process probabilistic constraints so efficiently. Uncovering the source of this efficiency could provide valuable insights for both advancing LLM development and deepening our understanding of human cognitive mechanisms.

5.2 Enhanced lingering misinterpretations from locality bias and limited revision

While LLMs displayed patterns similar to humans, they produced more error responses in garden-path conditions than humans. This trend was consistently observed across all experiments, regardless of verb type, length, head position, or plausibility. Though trained on an overwhelming amount of input with considerable computational power, LLMs struggled to overcome misinterpretations caused by garden-path sentences, tending to be more error-prone than human processors. This might suggest that LLMs either possess certain characteristics that make them more susceptible to garden-path effects, or they lack certain human-like features, which help mitigate these errors.

Falling into a garden-path trap occurs when a processor forms a temporally or locally coherent structure during sequential processing. In this regard, the occurrence of the garden-path effect in LLMs indicates that they are effectively carrying out sequential processing and properly forming the locally coherent structure (in this experiment, the verb-direct object relationship). However, since LLMs are 'too' susceptible to garden-path effects, there may be inherent vulnerabilities in how LLMs engage in sequential processing and form locally coherent structures.

One possibility is that the unidirectional processing nature of GPTs could heighten their vulnerability to garden-path sentences. While RNNs process tokens sequentially and must wait for the previous token to be processed before moving on to the next (Rumelhart et al. 1986), Transformers such as GPTs process all tokens in a sequence in parallel, which is known to incredibly reduce the calculation load of the model (Vaswani et al. 2017). However, despite its parallel processing nature, the decoder part of Transformers is inherently designed to respect the sequential nature of language through the masked self-attention mechanism. This ensures that when calculating the representation for a token, the model only incorporates information from preceding tokens, thus preserving the sequential flow of contextual information. For example, when calculating the contextual information of the word "deer," the model only uses the information from the previous tokens, "While the man hunted the". In this way, Transformers include the mechanisms that allow it to mimic human-like sequential processing within a parallel processing framework.

In particular, GPTs differ from BERT and other Transformers by exclusively using the decoder architecture with unidirectional processing. While Transformer encoders allow bidirectional processing (Vaswani et al. 2017), GPT uses only the decoder, which cannot utilize future context in word prediction. This makes GPT processing more strictly sequential than human sentence processing. Although human processing is generally sequential, eye-movement studies reveal more complex patterns involving quick movements, brief stops, occasional text skipping, and regressions to earlier sections (Rayner et al. 2005). Compared to this flexible human approach, GPT's strictly sequential processing may make it more susceptible to garden-path traps.

Another possibility is LLMs' greater reliance on locality compared to humans. The learning method of LLMs, based on next-word prediction, inherently depends heavily on the immediate locality, the relationship with nearby words, since the most crucial factor in predicting the next word is typically the preceding word. Traditional RNN-based language models were known for being heavily influenced by immediate past context, which has been highlighted as a significant drawback (Sherstinsky 2020). Although Transformers aimed to overcome this limitation through the attention mechanism that consider context from all words within a sentence (Vaswani et al. 2017), this locality bias appears to persist. Clark et al. (2019) indicated that even BERT's attention layers still place significant weight on the information from initial words in processing sentences. The results of this experiment could suggest that the dependency on locality remains strong in GPTs as well, and its impact may be more pronounced compared to humans.

Even if LLMs are more prone to garden-path effects, error responses in these experiments would not have emerged if they could effectively recover from such misinterpretations. In human sentence processing, initial misinterpretations following garden-path effects can impact global comprehension when the first misanalysis fails to be completely removed (Slattery et al. 2013). LLMs may be even more susceptible to such lingering misinterpretations. In the parallel processing system of the attention mechanism, early word processing not only influences subsequent word processing but also persists largely unchanged through to the final layer. Due to this mechanism, a syntactically coherent structure formed from earlier contexts is nearly unchanged when it reaches the final interpretation. Therefore, compared to humans, LLMs might be more susceptible to such lingering misinterpretations.

Moreover, the absence of a "revision" process in LLM can exacerbate this issue. In human sentence processing, awareness of ambiguity or temporary ungrammaticality typically arises upon reaching the main verb (here, "ran"), which triggers reanalysis, evidenced by increased reading times or eye-movement regressions (Frazier and Rayner 1982; for opposing views, see Christianson et al. 2017). While LLMs might be capable of recognizing temporary ungrammaticality in the disambiguation area, their ability to revise such errors is likely limited due to their operational mechanisms. Particularly, in the case of GPT, due to its unidirectional processing, the model lacks information from subsequent words, meaning it cannot correct an already formed incorrect relationship based on later word awareness. Thus, GPT and similar LLMs are incapable of performing a "revision" process, which could make them more susceptible to lingering misinterpretations.

If the absence of revision significantly contributes to the higher error rates observed in LLMs, it may be beneficial to consider designing a mechanism for reanalysis within LLMs. As noted, LLMs might be capable of detecting temporary ungrammaticalities, and these detections can propagate through subsequent layers, making structural reanalysis feasible. For example, BERT, being bidirectional, allows computations that occur later in the sentence to influence earlier words, incorporating context from both before and after a word in its predictions (Devlin et al. 2019). However, current evidence suggests that BERT, despite its bidirectional processing, also remains error-prone and does not consistently outperform humans in handling the garden-path sentences (Irwin et al. 2023). It appears that merely having bidirectional processing does not guarantee actual revision. To effectively facilitate revision, a direct structural addition might be necessary to explicitly trigger such processes. In fact, Madureira et al. (2024) attempted to design a model that enables such revisions, and their model demonstrated an improved capabilities in handling revisions.

5.3 Model comparison and the adaptive value of good-enough processing

Another focus of this study was the comparison of syntactic processing between GPT-3.5 and GPT-4. With the transition from GPT-3.5 to GPT-4, there has been a notable enhancement in overall language capabilities. The question is whether this enhanced ability extends to syntactic processing as well.

The results indicate that advancements in the models have contributed to better syntactic processing capabilities in specific conditions, though not universally. While the syntactic processing patterns of GPT-3.5 and GPT-4 were generally similar, GPT-4 displayed more human-like patterns and committed fewer errors in certain contexts. In the processing of unambiguous sentences, GPT-4 consistently outperformed GPT-3.5 across all tested experiments, irrespective of the disambiguation methods. However, in ambiguous garden-path conditions, GPT-4 did not show clear advantages over GPT-3.5. Both models exhibited similarly high error rates in these challenging conditions. Furthermore, in some highly ambiguous conditions such as plausible long sentences with OT verbs, GPT-4 actually produced more errors than GPT-3.5, with error rates approaching 100%. In sum, improvements from GPT-3.5 to GPT-4 seem to benefit syntactic processing selectively rather than comprehensively.

Understanding why these improvements were selective rather than comprehensive requires examining what changed between the models. Pinpointing the precise factors is challenging, as not all technical details about GPT-4 are publicly available. However, it can be inferred that the most substantial improvements likely stem from increases in model size and the amount of training data. Since GPT-2, its successors have employed the same mechanism of next-word prediction (Achiam et al. 2023) and have primarily used the decoder part of the Transformers (Brown et al. 2020). The scale of the models and the volume of training data, however, have expanded exponentially (Brown et al. 2020). For example, GPT-2 was equipped with 1.5 billion parameters (Radford et al. 2019), GPT-3 saw a dramatic increase to 175 billion parameters (Brown et al. 2020), and GPT-4 is estimated to exceed 1 trillion parameters (Islam and Moushi 2025). According to the scaling laws (Kaplan et al. 2020), the performance of LLMs scales as a power law with the size of the model and the dataset, while other architectural details have relatively minimal effects. Based on this, our discussion will proceed with the assumption that increases in model size or training data have been the most critical factors driving the advancements, at least according to the most documented and theoretically supported factors.[1]

These changes in model scale and training data also lead to improvements in syntactic processing, but as demonstrated in our experiments, these improvements were selective rather than comprehensive. This pattern might reveal insights about the limitations of scaling. While larger models brought advances in processing unambiguous sentences, they showed persistent difficulties in garden-path conditions, suggesting that some aspects of syntactic processing may not improve simply through

[1] Nevertheless, as one reviewer pointed out, performance improvements cannot be attributed solely to "larger models and more data." Although the basic architecture has remained consistent, other factors may have also contributed to performance gains, including architectural refinements (e.g., modifications to attention mechanisms or normalization techniques), optimization algorithm enhancements, advances in post-training methods such as RLHF (Reinforcement Learning from Human Feedback; Christiano et al. 2017; Ouyang et al. 2022), and improvements in training data quality and curation (Gunasekar et al. 2023). Given the lack of comprehensive technical documentation for GPT-4 (Achiam et al. 2023), our discussion focuses primarily on the most documented and theoretically supported factors (scale and data) while acknowledging that the actual improvements likely result from a combination of multiple technical innovations.

increased scale. The scaling law itself indicates that performance gains are not infinite. As models grow larger, the efficiency of these gains diminishes, following a power law relationship rather than linear improvement (Kaplan et al. 2020). This suggests that continued expansion beyond GPT-4's scale may yield progressively smaller returns, including for the challenging syntactic structures that have shown some resistance to improvement.

Furthermore, increased model size may actually worsen performance in certain contexts. Our experiments revealed that GPT-4 exhibited a paradoxical pattern of nearly binary responses, with error rates approaching either 0% or 100%. In highly ambiguous conditions, particularly sentences combining plausible interpretations, longer ambiguous regions, and OT verbs, GPT-4 produced error rates near 100%, performing worse than GPT-3.5. These high-error conditions share a common feature in that they increase the probabilistic likelihood of forming verb-object relationships. Conversely, GPT-4 demonstrated nearly perfect accuracy in minimally ambiguous conditions where the probabilistic likelihood of forming incorrect verb-object relationships was substantially lower. Combining these findings, it appears that GPT-4 demonstrated near-maximal errors when probabilistic patterns strongly favored garden-path interpretations and near-perfect accuracy when such patterns were unlikely.

This binary pattern might suggest a form of over-confidence stemming from excessive training data. Broadly speaking from a statistical standpoint, increasing the number of observations typically enhances the reliability of the results, reducing sampling error and providing a more precise reflection of the entire population. This larger sample size consequently narrows the confidence interval, enabling models to specify with greater certainty the range within which the true population parameter lies. However, a complication arises because language is constantly evolving and there is no definitive 'complete population' for it. In this context, LLMs might be learning from an excessively large input or a huge sample size, which could lead to an undue level of "confidence" in their estimations. From a machine learning perspective, this could represent a form of overfitting that may not manifest in typical ways during training but could lead the models to be overly confident in their language processing, when facing novel or ambiguous language data.

As observed in the results of this experiment, this type of overfitting might lead to nearly error-free parsing, seemingly providing an advantage in syntactic processing.

However, it also rendered GPT-4 more prone to errors compared to GPT-3.5, with error rates approaching 100% under several conditions. For instance, plausible long sentences with OT verbs typically introduce temporary ambiguity, but GPT-4 processed these as if there was no ambiguity, likely due to excessive "confidence" in the predicted verb-noun phrase relationship. In this regard, too much training data can indeed be detrimental in scenarios where sentence structures, while appearing highly probable within the constraints of the trained data or existing language data, may not be appropriate within the context of newly generated utterances.

Moreover, LLM's nearly error-free parsing shown in non-garden-path sentences might not always be an advantage, given the inherent unpredictability of human language, which often includes inaccurate or syntactically incomplete usage (Ferreira and Patson 2007). In real-world applications, a degree of flexibility in understanding and interpreting language could be more beneficial than presuming '100 percent' correctness, as it allows the models to adapt better to the diverse and evolving nature of human communication. Additionally, presuming '100 percent' correctness might inhibit learning possibilities. If an assumption deemed absolutely correct is later proven wrong, a significant weight shift is required, making the process of updating probabilities more challenging. In humans, a high subjective confidence rating has been reported to hinder word learning by preventing the acceptance of counter evidence (Dautriche et al. 2021). While current LLMs are not adaptive learners that continuously update as with humans, there is a substantial possibility that this could act as a restrictive factor.

GPT-4's over-confidence paradoxically reveals why human good-enough processing exists as an adaptive strategy. While good-enough processing in humans might be interpreted as a limitation stemming from cognitive constraints, GPT-4's behavior suggests that maintaining moderate confidence levels may be functionally advantageous rather than simply suboptimal. GPT-4 may show what happens when probabilistic constraints become excessively strong. This suggests that some degree of incomplete reanalysis may reflect a processing strategy that appropriately balances accuracy with adaptability. From this perspective, the moderate confidence levels characteristic of human sentence processing may be better suited to real-world language use than either insufficient processing or the excessive certainty.

6. Conclusion

This study investigated how advanced LLMs such as GPT-3.5 and GPT-4 process garden-path sentences, comparing their syntactic processing to human capabilities to improve our understanding of both machine and human language processing. The findings reveal that LLMs exhibit similar "good-enough" syntactic processing to humans, including making erroneous responses. This similarity suggests that LLMs may mimic human processing patterns through their learned probabilistic constraints. This might offer support for the theory that human language acquisition and processing can be explained through statistical learning, although the vast amount of data used to train LLMs necessitates identifying a mechanism that explains the high efficiency of human processing. The study also found that LLMs make more errors under garden-path conditions than humans. It is possibly due to LLMs' inherent characteristics such as their unidirectional processing nature and greater reliance on immediate locality without the ability to revise their responses. Furthermore, GPT-4 showed selective improvements over GPT-3.5, performing better in unambiguous conditions but not in garden-path conditions. This suggests limitations to scaling, as increased model size led to overconfidence in highly ambiguous contexts. This overconfidence paradoxically demonstrates the adaptive advantage of human good-enough processing with moderate confidence levels. To conclude, this study aimed to explore the underlying mechanisms of syntactic processing in LLMs using a targeted evaluation approach, comparing these mechanisms to those in humans. While the study relies on superficial outcomes due to the unavailability of the code for LLMs, this method is still meaningful as it provides insights that can contribute to our understanding of both human and machine language processing. Future research should aim to unveil the yet undiscovered mechanisms of LLMs through more sophisticated paradigms, thereby enhancing our understanding of both machine and human language.

References

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida et al. 2023. GPT-4 technical report. *arXiv preprint* arXiv:2303.08774.

Altmann, Gerry T. M. 2013. Anticipating the garden path: The horse raced past the barn ate the cake. In Montserrat Sanz, Itziar Laka, and Michael K. Tanenhaus (eds.), *Language down the garden path: The cognitive and biological basis for linguistic structure*, 111-130. Oxford: Oxford University Press.

Baayen, R. Harald, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4): 390-412.

Bacon, Geoff and Terry Regier. 2019. Does BERT agree? Evaluating knowledge of structure dependence through agreement relations. *arXiv preprint* arXiv:1908.09892.

Bazhukov, Maxim, Ekaterina Voloshina, Sergey Pletenev, Arseny Anisimov, Oleg Serikov, and Svetlana Toldova. 2024. Of models and men: Probing neural networks for agreement attraction with psycholinguistic data. In Libby Barak and Malihe Alikhani (eds.), *Proceedings of the 28th Conference on Computational Natural Language Learning*, 280-290. Miami: Association for Computational Linguistics.

Bever, Thomas G. 1970. The cognitive basis for linguistic structures. In John R. Hayes (ed.), *Cognition and the development of language,* 279-362. New York: Wiley and Sons.

Bojic, Ljubisa, Predrag Kovacevic, and Milan Cabarkapa. 2023. GPT-4 surpassing human performance in linguistic pragmatics. *arXiv preprint* arXiv:2312.09545.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33: 1877-1901.

Cai, Zhenguang, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In Tatsuki Kuribayashi, Giulia Rambelli, Ece Takmaz, Philipp Wicke, and Yohei Oseki (eds.) *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 37-56. Bangkok: Association for Computational Linguistics.

Chaves, Rui P. 2020. What don't RNN language models learn about filler-gap dependencies? *Proceedings of the Society for Computation in Linguistics* 3(1): 20-30.

Christiano, Paul, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems* 30: 4299-4307.

Christianson, Kiel, Andrew Hollingworth, John F. Halliwell, and Fernanda Ferreira. 2001. Thematic roles assigned along the garden path linger. *Cognitive Psychology 42*(4): 368-407.

Christianson, Kiel., Steven G. Luke, Erika K. Hussey, and Kacey L. Wochna. 2017. Why reread? Evidence from garden-path and local coherence structures. *Quarterly Journal of Experimental*

Psychology 70(7): 1380-1405.

Clark, Kevin, Urvashi Khandelwal, Omer Levy and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes (eds.) *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276-286. Florence: Association for Computational Linguistics

Dautriche, Isabelle, Hugh Rabagliati and Kenny Smith. 2021. Subjective confidence influences word learning in a cross-situational statistical learning task. *Journal of Memory and Language* 121: 104277.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.) *Proceedings of the 2019 Conference of NAACL-HLT*, 4171-4186. Minneapolis: Association for Computational Linguistics.

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14(2): 179-211.

Ettinger, Allyson. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics* 8: 34-48.

Frank, Stefan and John C. J. Hoeks. 2019. The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 41: 337-343.

Frazier, Lyn and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14(2): 178-210.

Ferreira, Fernanda, and Nikole D. Patson. 2007. The 'good enough' approach to language comprehension. *Language and Linguistics Compass* 1(1-2): 71-83.

Futrell, Richard, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint* arXiv:1809.01329.

Futrell, Richard, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.) *Proceedings of the 2019 Conference of NAACL-HLT*, 32-42. Minneapolis: Association for Computational Linguistics.

Goldberg, Yoav. 2019. Assessing BERT's syntactic abilities. arXiv preprint arXiv:1901.05287.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.) P*roceedings of the 2018 Conference of NAACL-HLT*, 1195-1205. New Orleans: Association for Computational Linguistics.

Gunasekar, Suriya, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil

Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *arXiv preprint* arXiv:2306.11644.

Gweon, Hyowon, Joshua. B. Tenenbaum, and Laura. E. Schulz. 2010. Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences* 107(20): 9066-9071.

Hauser, Marc D., Noam Chomsky, and W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298(5598): 1569-1579.

Herbold, Steffen, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva and Alexander Trautsch. 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports* 13: 18617.

Hu, Jennifer, Jon Gauthier, Peng Qian, Ethan Wilcox and Roger P. Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 1725-1744. Online: Association for Computational Linguistics

Hunter, John D. 2007. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering* 9(03): 90-95.

Irwin, Tovah, Kyra Wilson, and Alec Marantz. 2023. BERT shows garden path effects. In Andreas Vlachos and Isabelle Augenstein (eds.) *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3220-3232. Dubrovnik: Association for Computational Linguistics.

Islam, R. and O. M. Moushi. 2025. GPT-4o: The cutting-edge advancement in multimodal LLM. In Kohei Arai (eds.) *Proceedings of the 2025 Computing Conference*, 47-60. London: Springer

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint* arXiv:2001.08361.

Kuncoro, Adhiguna, Lingpeng Kong, Daniel Fried, Dani Yogatama, Laura Rimell, Chris Dyer, and Phil Blunsom. 2020. Syntactic structure distillation pretraining for bidirectional encoders. *Transactions of the Association for Computational Linguistics* 8: 776-794.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune HB Christensen. 2017. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* 82: 1-26.

Lee, Jonghyun and Jeong-Ah Shin. 2023. Decoding BERT's internal processing of garden-path structures through attention maps. *Korean Journal of English Language and Linguistics* 23: 461-481.

Lee, Jonghyun, Jeong-Ah Shin, and Myung-Kwan Park. 2022. (AL)BERT down the garden path: Psycholinguistic experiments for pre-trained language models. *Korean Journal of English*

Language and Linguistics 22: 1033-1050.

Lenth, Russell. 2023. Emmeans: Estimated marginal means, aka least-squares means. R package version 1.8.6.

Linzen, Tal and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics* 7: 195-212.

Linzen, Tal, 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.

MacDonald, Maryellen C. 2013. Inviting production to the cognitive basis party. In Montserrat Sanz, Itziar Laka, and Michael K. Tanenhaus (eds.), *Language down the garden path: The cognitive and biological basis for linguistic structure*, 131-140. Oxford: Oxford University Press.

Madureira, Brielen, Patrick Kahardipraja, and David Schlangen. 2024. When only time will tell: Interpreting how transformers process local ambiguities through the lens of re-start-incrementality. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* 1, 4722-4749. Bankok: Association for Computational Linguistics.

Marvin, Rebecca and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.) *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* 1192-1202. Brussels: Association for Computational Linguistics.

McCoy, R. Thomas, Robert Frank and Tal Linzen. 2020. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics* 8. 125-140.

Orru, Graziella, Andrea Piarulli, Ciro Conversano, and Angelo Gemignani. 2023. Human-like problem-solving abilities in large language models using ChatGPT. *Frontiers in Artificial Intelligence* 6: 1199350.

Ouyang, Long, J. Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35: 27730-27744.

Qiu, Zhuang, Xufeng Duan, and Zhenguang G. Cai 2025. Grammaticality representation in ChatGPT as compared to linguists and laypeople. *Humanities and Social Sciences Communications* 12(1): 1-15.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1(8): 9.

Rayner, Keith, Barbara J. Juhasz and Alexander Pollatsek. 2005. Eye movements during reading. In Margaret J. Snowling and Charles Hulme (eds.), *The science of reading: A handbook*, 79-97. Oxford: Blackwell Publishing.

Rumelhart, David E., Geoffrey E. Hinton and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323: 533-536.

Saffran, Jenny. R., Richard N. Aslin and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274(5294): 1926-1928.

Salverda, Anne Pier, Meredith Brown and Michael K. Tanenhaus. 2011. A goal-based perspective on eye movements in visual world studies. *Acta Psychologica* 137(2). 172-180.

Sherstinsky, Alex. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena* 404: 132306.

Slattery, Timothy J., Patrick Sturt, Kiel Christianson, Masaya Yoshida, and Fernanda Ferreira. 2013. Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language* 69(2): 104-120.

Snowling, Margaret J., Charles Hulme and Kate Nation (eds.). 2022. *The science of reading: A handbook.* Hoboken: John Wiley and Sons.

Suresh, Siddharth, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, and Timothy T. Rogers. 2023. Conceptual structure coheres in human cognition but not in large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* 722-738. Singapore: Association for Computational Linguistics.

Taloni, Andrea, Massimiliano Borselli, Valentina Scarsi, Costanza Rossi, Giulia Coco, Vincenzo Scorcia, and Giuseppe Giannaccare. 2023. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. *Scientific Reports* 13: 18562.

van Schijndel, Marten, Aaron Mueller and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In Kentaro Inui, Jing Jiang, aVincent Ng, and Xiaojun Wan (eds.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5831-5837. Hong Kong: Association for Computational Linguistics.

van Schijndel, Marten and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 40: 2603-2608.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30: 5999-6009.

Wang, Andrew Y., Sherman Lin, Christopher Tran, Robert J. Homer, Dan Wilsdon, Joanna C. Walsh, Emily A. Goebel, Irene Sansano, Snehal Sonawane, Vincent Cockenpot, Sanjay Mukhopadhyay, Toros Taskin, Nusrat Zahra, Luca Cima, Orhan Semerci, Birsen Gizem Özamrak, Pallavi Mishra, Naga Sarika Vennavalli, Po-Hsuan Cameron Chen, and Matthew

J. Cecchini. 2024. Assessment of pathology domain-specific knowledge of ChatGPT and comparison to human performance. *Archives of Pathology and Laboratory Medicine* 148(10): 1152-1158.

Waskom, Michael L. 2021. Seaborn: statistical data visualization. *Journal of Open Source Software* 6(60): 3021.

Wilcox, Ethan, Roger Levy and Richard Futrell. 2019. Hierarchical representation in neural language models: Suppression and recovery of expectations. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes (eds.) *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 181-190. Florence: Association for Computational Linguistics.

Wilcox, Ethan, Roger Levy, Takashi Morita and Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.) *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 211-221. Brussels: Association for Computational Linguistics.

Zhai, Xiaoming, Matthew Nyaaba, and Wenchao Ma. 2024. Can generative AI and ChatGPT outperform humans on cognitive-demanding problem-solving tasks in science? *Science and Education* 1-22.

Jonghyun Lee
Assistant Professor
English Studies Major, Divison of Global Studies, College of Global Business
Korea University Sejong Campus
2511 Sejong-ro
Sejong, 30019, Korea
E-mail: j-lee@korea.ac.kr

Jeong-Ah Shin
Professor
Department of English Language and Literature
Dongguk University
30 Pildong-ro 1-gil, Jung-gu
Seoul, 04620, Korea
E-mail: jashin@dongguk.edu

Received: 2025. 08. 29.
Revised: 2025. 10. 27.
Accepted: 2025. 11. 14.