



## Probing Good-Enough Processing in Large Language Models with a Paraphrasing Task\*

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



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

Received: October 10, 2025

Revised: January 12, 2026

Accepted: January 13, 2026

Lee, Jonghyun (First author)  
Assistant Professor, English  
Studies Major, Division of Global  
Studies  
Korea University Sejong Campus  
2511 Sejong-ro, Jochiwon-eup  
Sejong, Korea  
Email: [j-lee@korea.ac.kr](mailto:j-lee@korea.ac.kr)

Shin, Jeong-Ah (Corresponding  
author)  
Professor, Department of English  
Language and Literature  
Dongguk University  
30, Pildong-ro 1-gil, Jung-gu  
Seoul, Korea  
Email: [jashin@dgu.ac.kr](mailto:jashin@dgu.ac.kr)

\* 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).

### ABSTRACT

Lee, Jonghyun and Jeong-Ah Shin. 2026. Probing good-enough processing in large language models with a paraphrasing task. *Korean Journal of English Language and Linguistics* 26, 127-141.

This study investigates whether large language models (LLMs) exhibit human-like ‘good-enough’ processing patterns in syntactic comprehension or demonstrate mechanical accuracy. Previous research using forced-choice question-answering paradigms revealed that LLMs display incomplete syntactic reanalysis similar to humans when processing garden-path sentences. However, concerns arose that these patterns might reflect methodological artifacts rather than genuine processing characteristics, as direct questioning could bias models toward initial misinterpretations. To address this limitation, we employed a paraphrasing task that requires comprehensive sentence reformulation rather than binary responses, following Patson et al. (2009). We tested GPT-3.5 and GPT-4 on 24 garden-path sentences containing Optionally Transitive (OT) and Reflexive Absolute Transitive (RAT) verbs. Results demonstrate that good-enough processing patterns persist across both paradigms, with LLMs continuing to exhibit partial reanalysis in garden-path conditions even when generating full paraphrases. This confirms that previously observed error patterns represent genuine syntactic processing characteristics rather than experimental artifacts. Notably, GPT-4 showed improved performance in the paraphrasing task compared to forced-choice experiments, suggesting task-dependent variation in processing depth. Both models exhibited human-like incomplete processing despite their substantial computational resources, indicating that their pattern-matching mechanisms favor processing shortcuts over complete syntactic interpretation. These findings reveal that LLMs demonstrate good-enough processing similar to humans, with performance varying systematically across task formats.

### KEYWORDS

large language models, garden-path sentences, good-enough processing, syntactic processing, paraphrasing task, ChatGPT

## 1. Introduction

The recent performance of Large Language Models (LLMs) has transformed long-standing skepticism about machine language understanding capabilities. For decades, researchers argued that machines could never achieve human-like language comprehension (Bender et al. 2021, Davis and Marcus 2015, Dreyfus 1972, Searle 1980). Yet current LLMs now demonstrate capabilities that seemed impossible just years ago (Bubeck et al. 2023), achieving human-level performance in reading comprehension (Shultz et al. 2025), natural language inference (Madaan et al. 2025), and text generation (Tian et al. 2024). Despite these advances across various language tasks, the mechanisms underlying such exceptional language processing remain largely unknown. This opacity is particularly pronounced in syntactic processing, long considered a uniquely human linguistic capability (Hauser et al. 2002). It remains unclear whether these computational operations truly reflect human-like syntactic processing and what underlying mechanisms enable such performance.

To investigate LLM syntactic abilities, researchers have adopted a targeted evaluation approach for systematic assessment (Linzen et al. 2016, Marvin and Linzen 2018). This methodology presents carefully constructed sentences that require specific syntactic representations for correct interpretation, following experimental techniques used in psycholinguistic research. Previous investigations using this approach have demonstrated that LLMs capture a variety of syntactic phenomena, including agreement patterns (Bacon and Regier 2019, Bernardy and Lappin 2017, Goldberg 2019, Gulordava et al. 2018, Hu et al. 2020, Linzen and Leonard 2018), filler-gap dependencies (Chaves 2020, Chowdhury and Zamparelli 2018, Futrell et al. 2018, Hu et al. 2020, Wilcox et al. 2018, Wilcox et al. 2019), reflexive anaphora (Goldberg 2019, Hu et al. 2020), negative polarity items (Futrell et al. 2018, Marvin and Linzen 2018), center-embedding structures (Futrell et al. 2018, Hu et al. 2020, Wilcox et al. 2019) and syntactic ambiguity such as garden-path sentences (Futrell et al. 2018, Hu et al. 2020, Huang et al. 2024, Lee and Shin 2023, Lee et al. 2022, van Schijndel and Linzen 2018, Wilcox et al. 2019). The targeted evaluation framework provides controlled testing of specific linguistic phenomena and enables direct comparison with human language processing patterns. This study utilizes these advantages of the targeted evaluation approach to investigate the syntactic processing capabilities of advanced LLMs.

The present study focuses on garden-path sentences, which are temporarily ambiguous structures that initially guide processors toward incorrect interpretations before requiring reanalysis for proper comprehension (Bever 1970, Frazier and Rayner 1982). This structure provides several methodological advantages that make them suitable for syntactic evaluation. Unlike NP/VP ambiguities where multiple interpretations remain equally valid, garden-path sentences require identification of a singular syntactically correct resolution, enabling clear assessment of comprehension accuracy. 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). Additionally, garden-path sentences reflect the complex interplay of linguistic factors such as semantic, morphological, and phonological cues that influence sentence processing during reanalysis, allowing researchers to examine how models prioritize and utilize different types of linguistic information in syntactic processing (Futrell et al. 2018, Lee and Shin 2023, Lee et al. 2022, van Schijndel and Linzen 2018).

Human processing of garden-path sentences has revealed numerous aspects of human syntactic processing mechanisms (Ferreira and Clifton 1986, Frazier and Rayner 1982, Gibson 1998, Gibson 2000, Levy 2008, MacDonald et al. 1994, Osterhout and Holcomb 1992, Rayner and Frazier 1987, Trueswell et al. 1994). Of particular relevance to the current study is the phenomenon of ‘good-enough’ processing (Christianson et al.

2001, Ferreira et al. 2002, Frances 2024, Patson et al. 2009, Slattery et al. 2013), which demonstrates incomplete reanalysis patterns in human comprehension. For example, when reading “While Mary bathed the baby played in the crib,” most readers incorrectly affirm that “Mary bathed the baby” while also correctly recognizing that “the baby played in the crib” (Christianson et al. 2001). Drawing from such evidence, Ferreira and Patson (2007) developed the ‘good-enough’ framework, arguing that readers often settle for rapid, incomplete sentence interpretations rather than pursuing comprehensive syntactic analysis. This processing pattern can be interpreted either as evidence of limitations in human syntactic capabilities or alternatively as an adaptive strategy that emphasizes speed and cognitive efficiency over complete accuracy (Ferreira and Patson 2007).

Given these findings about human syntactic processing, the question arises as to whether LLMs might exhibit similar good-enough processing patterns or demonstrate mechanical accuracy. On one hand, LLMs possess substantial computational resources and lack the cognitive limitations that constrain human processing (MacDonald et al. 1992), including working memory restrictions, limited attentional capacity, and processing speed constraints. This could potentially enable more complete syntactic analysis, which would entail constructing a fully specified, globally consistent parse that incorporates all available syntactic constraints from the input. On the other hand, LLMs might display human-like good-enough processing because they learn from vast amounts of human-produced text that already contains the patterns of good-enough processing, potentially inheriting these patterns through their training data. Moreover, since LLM processing fundamentally operates through probabilistic pattern-matching, they may naturally develop similar processing shortcuts by favoring the most frequent linguistic patterns over computationally expensive reanalysis. This study explores which of these possibilities characterizes LLM garden-path processing by examining their syntactic behavior in controlled experimental conditions.

Investigations into LLM garden-path processing have revealed evidence of good-enough processing characteristics similar to those observed in humans (Lee and Shin 2025). Previous research has demonstrated that both GPT-3.5 and GPT-4 exhibit the characteristic pattern of good-enough processing: successful structural reanalysis of main clause subjects while simultaneously maintaining incorrect subordinate clause interpretations. However, LLMs display an exaggerated version of this phenomenon, consistently producing higher error rates than human processors in garden-path conditions. This tendency appeared similarly in both models. Although GPT-4 generally demonstrated more human-like patterns and achieved higher overall accuracy in non-garden-path sentences, it exhibited similar levels of error response patterns to GPT-3.5 in garden-path conditions and sometimes produced even more errors in specific conditions where initial misinterpretations were highly likely due to plausibility and longer sentence length. These findings indicate that LLMs exhibit incomplete syntactic processing similar to humans, with even the more advanced model displaying the same good-enough processing patterns.

However, these findings have limitations due to their use of forced-choice question-answering paradigms that may bias responses toward initial misinterpretations (Patson et al. 2009). In the previous research, models were presented with sentences such as “While the man hunted the deer ran into the woods” and then asked direct questions such as “Did the man hunt the deer?” requiring simple yes/no responses, following the methodology established by Christianson et al. (2001), which first demonstrated good-enough processing in humans using this yes/no question paradigm. However, Patson et al. (2009) later raised methodological concerns about this approach, suggesting that direct questioning could predispose participants toward their initial misinterpretation. Due to its imperfect nature, human memory can reconstruct memories based on given cues and trigger non-existent memories through questioning (Loftus and Pickrell 1995). This methodological concern is relevant for LLMs as well, which is susceptible to question-induced hallucination or adaptive response generation without

reflecting genuine comprehension (Ardoin et al. 2025, Farquhar et al. 2024, Sharma et al. 2024, Turpin et al. 2023). The conversational nature of modern LLMs creates potential for response adaptation based on perceived question expectations rather than authentic linguistic analysis. Given these concerns, the good-enough processing patterns observed in previous LLM experiments may not reflect authentic syntactic processing but could instead be artifacts of these methodological limitations.

To circumvent the possible biases of question-answering methods and more directly reveal persistent misinterpretations, Patson et al. (2009) introduced a paraphrasing task. In this task, participants read garden-path sentences and subsequently paraphrased them, compelling them to formulate and express their ultimate interpretation of the sentences. This technique is designed to more accurately capture the reader's final interpretation by engaging them in free recall, reducing the likelihood of influencing them to default to the original, possibly incorrect interpretation. The current study adapts this paraphrasing paradigm from Patson et al. (2009), which requires models to generate comprehensive sentence reformulations rather than simple binary responses. This may reveal more authentic patterns of syntactic interpretation while minimizing experimental demand characteristics that could compromise the validity of findings about LLM language processing capabilities.

In this regard, the present study investigates whether the good-enough processing patterns observed in previous research persist when methodological biases are minimized through a paraphrasing paradigm. Following Patson et al. (2009), this study employs a paraphrasing task that requires GPT-3.5 and GPT-4 to reformulate garden-path sentences, thereby revealing their syntactic interpretations without the influence of direct questioning.

The current study examines sentences with two verb types, Optionally Transitive (OT) and Reflexive Absolute Transitive (RAT) verbs, across garden-path and non-garden-path conditions. OT verbs such as *hunt*, *sail*, and *read* can optionally take a direct object. When followed by a noun phrase, these verbs create temporary ambiguity about whether the subsequent noun serves as the verb's object, potentially leading to garden-path misinterpretation in subordinate clause structures. RAT verbs, typically associated with personal hygiene activities such as *wash*, *bathe*, and *shave*, similarly can optionally take a direct object and thus induce garden-path effects. However, RAT verbs differ from OT verbs in that they are grammatically required to be interpreted reflexively when no direct object is present. For instance, "Mary bathed" must be understood as "Mary bathed herself," whereas "Mary hunted" does not carry this reflexive interpretation.

This grammatical property prevents the pragmatic inference that can occur with OT verbs. In "While Mary hunted the deer ran into the woods," readers might interpret "the deer" as the object of "hunted" through pragmatic reasoning that the unspecified object could plausibly be "the deer." However, in "While Anna bathed the baby spit up on the bed," if readers properly reanalyze the sentence and recognize the grammatical properties of RAT verbs, "Anna bathed" should be interpreted as "Anna bathed herself," thereby preventing the pragmatic misinterpretation that Anna bathed the baby.

The inclusion of these two verb types allows for examining the extent to which pragmatic inference versus incomplete syntactic reanalysis contributes to misinterpretations. If errors are largely driven by pragmatic reasoning, error rates with RAT verbs should be substantially reduced in garden-path conditions, as these verbs block pragmatic inference. Previous human studies (Christianson et al. 2001, Patson et al. 2009) showed that garden-path effects persisted with both OT and RAT verbs. Although in both studies, RAT verbs yielded lower error rates than OT verbs, participants still produced considerable errors with RAT verbs in garden-path conditions despite these verbs blocking pragmatic inference. This suggests that misinterpretations in garden-path sentences stem largely from incomplete syntactic reanalysis rather than from pragmatic reasoning alone.

Following the approach established by the human studies, the present study uses these two verb types to determine the source of LLMs' potential errors, whether incomplete syntactic reanalysis or pragmatic reasoning.

This design addresses three primary research questions: First, do LLMs continue to exhibit good-enough processing characteristics when required to generate comprehensive paraphrases rather than binary responses? Second, if LLMs demonstrate good-enough processing patterns, do these errors stem from incomplete syntactic reanalysis or from pragmatic reasoning? Third, does the more advanced GPT-4 demonstrate more complete syntactic reanalysis compared to GPT-3.5, or do both models exhibit similar patterns of incomplete processing? By examining these questions, we aim to determine whether previously observed good-enough processing patterns in LLMs reflect genuine syntactic processing mechanisms or merely artifacts of question-answering methodologies.

## 2. Method<sup>1</sup>

### 2.1 Models

This research employed two large language models from OpenAI's GPT family: GPT-3.5 (gpt-3.5-turbo-0613) and GPT-4 (gpt-4-0613). These models represent successive generations in OpenAI's development trajectory, with GPT-4 constituting an architectural advancement beyond its predecessor. GPT-4 exhibits enhanced capabilities across multiple dimensions of language processing compared to GPT-3.5. The technical specifications detailed indicate improvements in several core areas: deeper contextual understanding, extended capacity for maintaining discourse coherence across lengthy text sequences, and heightened precision in generating contextually appropriate responses (Achiam et al. 2023). Additionally, GPT-4 demonstrates strengthened performance on tasks demanding complex reasoning and analytical problem-solving, stemming from both its enlarged training corpus and refined inferential mechanisms. By examining both models in parallel, this study investigates whether advances in general language modeling capabilities translate into corresponding improvements in syntactic processing performance, particularly when handling garden-path structures.

### 2.2 Procedure

Model testing was conducted through OpenAI's chat completion API. During the task, the models received instructions to paraphrase the sentences they read, avoiding mere repetition and encouraged to employ synonyms (e.g., "Paraphrase the sentence. But do not simply repeat the sentence. It is recommended to use synonyms."). Each sentence was tested ten times, with results averaged across trials. All trials used zero-shot prompts with no prior context or examples. Trials were fully independent. The conversation context was reset after each response, ensuring that prior trials could not influence subsequent responses.

---

<sup>1</sup> The complete code and dataset for this study can be accessed at:  
<https://github.com/coolmintmild/GoodEnoughParaphrasing>.

## 2.3 Materials

The experiment utilized 24 sentences identical to those used in Patson et al. (2009), which were originally from Christianson et al. (2001). This allowed for comparison between LLM and human data. Of these, 12 sentences contained OT verbs such as in (1), while the remaining 12 employed RAT verbs such as in (2).

- 1) 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 – Non-garden-path]
- 2) 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 – Non-garden-path]

Garden-path sentences (1a, 2a) and their non-garden-path counterparts (1b, 2b) differed minimally, distinguished only by the presence of a comma after the subordinate clause verb. This comma functions as a syntactic boundary marker that prevents garden-path misinterpretation by clearly demarcating the clause structure. This minimal manipulation allows for controlled comparison while isolating the effect of structural ambiguity.

## 2.4 Scoring

Paraphrase responses were classified into four distinct categories: failed, partial, full reanalysis, and others. Failed reanalysis occurred when the model retained its initial incorrect interpretation, failing to recognize the subject of the main clause accurately (for instance, “The man hunted the deer.”). Partial reanalysis was recognized when the model ambiguously interpreted the noun phrase “the deer” as both the direct object of the subordinate clause verb “hunted” and the subject of the main clause “ran” (e.g., “The man hunted the deer and it ran into the woods.”). Full reanalysis indicated that the model correctly identified the noun phrase “the deer” as the subject of the main clause while effectively dissociating it from the subordinate clause verb (e.g., “The man hunted and the deer ran into the woods.”). Others fall into a category that does not fit any of the predefined classifications, as the model produced alternative interpretations that may not necessarily be accurate.

## 2.5 Statistical Analysis

The statistical analysis was conducted in two stages. Initially, a Chi-Squared Test of Independence was utilized to determine if there were significant differences in the proportions between the two models. This was performed in the form of a stratified analysis, executing separate Chi-squared tests for (1) the entire data set, (2) each level of Garden-path (garden-path and non-garden-path), (3) each level of Verb Type (OT and RAT), and (4) each combination of Garden-path and Verb Type (garden-path OT, non-garden-path OT, garden-path RAT, non-garden-path RAT).

In the second stage, Generalized Linear Mixed-effect Model (GLMM) was employed to analyze the effects of LLMs, Garden-path, and Verb Type on the responses. To fit the GLMM, the dependent variables were converted

into binary data. Among the four paraphrase response categories, ‘Full reanalysis’ responses, which align with accurate syntactic interpretation of the sentences, were analogous to ‘correct responses’ from ‘yes/no’ binary questions and were encoded as 1. The other three response categories, deemed as erroneous, were encoded as 0. The model was structured in a 2×2×2 factorial design (RAT vs. OT × Garden-path vs. Non-garden-path × GPT-3.5 vs. GPT-4), with Verb Type, Garden-path, and Model as fixed effects and Items as random effects.

### 3. Results

Figure 1 shows that GPT-3.5 and GPT-4 both exhibited a higher frequency of full reanalysis in non-garden-path conditions than in garden-path conditions. Within the garden-path conditions, while both models predominantly produced partial reanalyses, GPT-4 demonstrated a lower count of partial and failed reanalyses compared to GPT-3.5. However, it is noteworthy that both models rarely produced failed analyses, indicating that they did not merely fail at syntactic reanalysis.

A chi-square test of independence was conducted to evaluate the response distribution between the two models. The test revealed a significant difference in response distribution between GPT-3.5 and GPT-4 across the entire dataset ( $\chi^2(3) = 33.857, p < 0.001$ ), in Garden-path conditions ( $\chi^2(3) = 50.606, p < 0.001$ ), with OT verbs ( $\chi^2(3) = 26.540, p < 0.001$ ), RAT verbs ( $\chi^2(3) = 9.585, p < 0.05$ ), and the combinations of Garden-path with OT ( $\chi^2(3) = 37.281, p < 0.001$ ) and with RAT verbs ( $\chi^2(3) = 18.016, p < 0.001$ ). However, no significant difference was found in all Non-garden-path conditions ( $p > 0.1$ ).

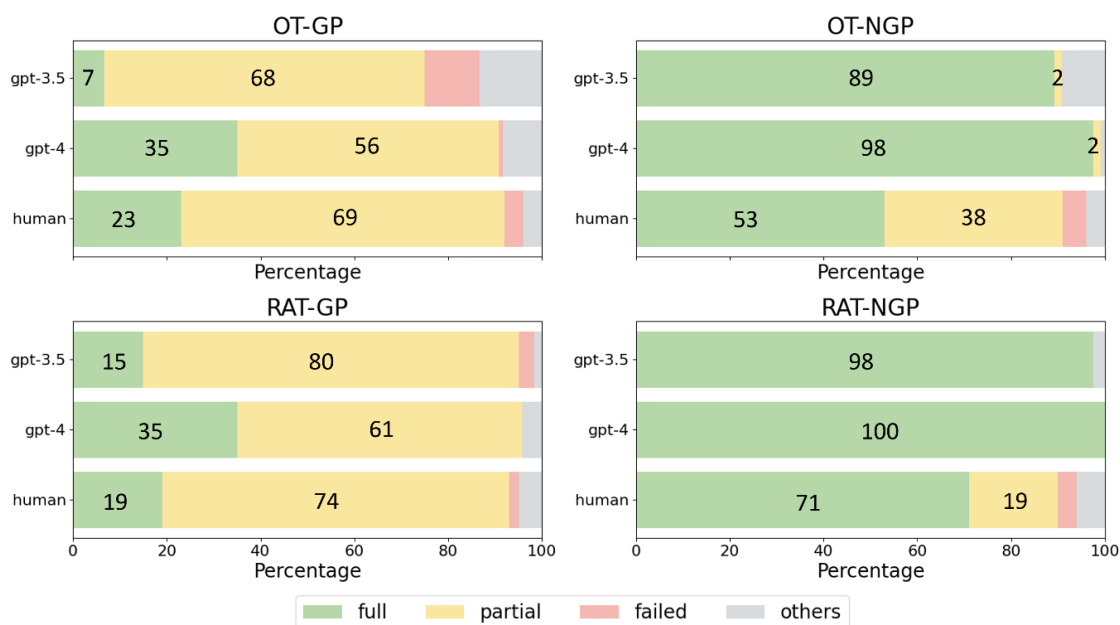


Figure 1. Percentage of Responses by Verb Type, Garden-path, and LMs

Note. The human data is sourced from Patson et al. (2009). OT=Optionally Transitive, RAT= Reflexive Absolute Transitive, GP=Garden-path, NGP=Non-garden-path.

The second phase of the analysis employed a GLMM to examine how each factor influenced responses and their interactions. The results revealed significant main effects for both Garden-path (estimate = -12.632, SE = 2.804,  $z = -4.506$ ,  $p < 0.001$ ) and LLMs (estimate = -4.857, SE = 0.928,  $z = -5.238$ ,  $p < 0.001$ ). Both GPT-3.5 and GPT-4 produced significantly more full analyses in Non-garden-path conditions compared to Garden-path conditions, and GPT-4 generated significantly more full analyses overall than GPT-3.5. A significant interaction between LLMs and Garden-path was also observed (estimate = 2.946, SE = 1.172,  $z = 2.514$ ,  $p < 0.05$ ). Post-hoc analysis revealed that GPT-4 produced significantly more full analyses than GPT-3.5 in Garden-path conditions (estimate = 3.840, SE = 0.567,  $z = 6.770$ ,  $p < 0.001$ ), while the two models showed no significant difference in Non-garden-path conditions ( $p > 0.1$ ).

In summary, both models produced more partial and failed analyses in Garden-path conditions compared to Non-garden-path conditions. While the response distributions were comparable between models in Non-garden-path conditions, GPT-4 outperformed GPT-3.5 in Garden-path conditions by generating a significantly higher proportion of full analyses.

Patson et al. (2009) observed that humans produced more partial reanalysis and less full reanalysis in garden-path conditions, indicating persistent misinterpretations in paraphrasing tasks as well as forced-choice paradigms. LLMs displayed similar patterns, showing reduced full reanalysis and increased partial reanalysis in garden-path conditions. However, in contrast to the forced-choice experiments reported in prior research, GPT-4 exhibited a higher accuracy rate than humans in Garden-path conditions when full reanalysis is treated as the correct response. Additionally, in Non-garden-path conditions, where human participants frequently showed partial analyses, both LLMs achieved nearly 100% full reanalysis.

## 4. Discussion

The current study investigated whether the good-enough processing patterns observed in prior research using forced-choice question-answering paradigms (Lee and Shin 2025) would persist when LLMs were required to paraphrase garden-path sentences. This methodological shift was designed to minimize potential biases inherent in direct questioning, which could predispose models toward initial misinterpretations (Patson et al. 2009). The findings revealed that LLMs continued to exhibit human-like syntactic patterns and this tendency was observed not only with OT verbs but also with RAT verbs. These results demonstrate that the error responses observed in prior research were not merely artifacts of the experimental paradigm but genuinely reflect lingering syntactic misinterpretations generated by LLMs.

The results align with prior forced-choice experiments in several respects. Both methodologies revealed that LLMs exhibit good-enough processing characteristics similar to humans (Patson et al. 2009), particularly the pattern of successful structural reanalysis of main clause subjects while simultaneously maintaining incorrect subordinate clause interpretations. In garden-path conditions, both GPT-3.5 and GPT-4 predominantly produced partial reanalyses, indicating that they recognized the correct subject of the main clause but failed to completely abandon the initial misinterpretation of the noun phrase as the object of the subordinate verb.

Moreover, the high error rates observed in the paraphrasing task were not solely attributable to pragmatic reasoning. The present study included both OT and RAT verbs specifically to examine the extent to which pragmatic inference contributed to lingering misinterpretations. Results showed that both GPT-3.5 and GPT-4 exhibited garden-path effects with both OT and RAT verbs at comparable error rates. This pattern suggests that the misinterpretations likely reflect incomplete syntactic reanalysis rather than arising solely from pragmatic

inference. However, LLMs may not fully recognize the obligatory reflexive properties of RAT verbs. If models fail to encode these verb-specific grammatical constraints, pragmatic inference could still operate where it should be grammatically blocked. Yet performance in non-garden-path conditions shows minimal errors with RAT verbs when no garden-path structure is present. This indicates that poor understanding of RAT verb properties cannot account for the observed error patterns. The evidence suggests that lingering misinterpretations largely reflect incomplete syntactic reanalysis.

The patterns observed in LLMs closely resemble those found in previous human studies. In both forced-choice question-answering (Christianson et al. 2001) and paraphrasing tasks (Patson et al. 2009), human participants demonstrated clear garden-path effects, producing more errors and partial reanalyses in garden-path conditions compared to non-garden-path conditions. Additionally, these patterns emerged regardless of verb type, indicating that such errors do not stem solely from pragmatic reasoning. The current study found similar patterns in LLMs, suggesting at least surface-level similarities in how LLMs and humans respond to temporarily ambiguous syntactic structures.

However, despite these surface-level similarities, differences between human and LLM performance were observed as well. One difference is that LLMs demonstrated substantially higher accuracy than humans in non-garden-path conditions across both yes/no and paraphrasing tasks. For example, human participants showed considerable error rates even in comma-disambiguated sentences, producing approximately 38% partial analyses for OT verbs and 19% for RAT verbs (Patson et al. 2009). In contrast, both GPT-3.5 and GPT-4 achieved nearly perfect performance in these conditions for both verb types. Human errors in non-garden-path sentences cannot be attributed to syntactic misanalysis since commas explicitly resolve structural ambiguity. While such errors may partly reflect memory limitations or simple miscomprehension, they likely stem largely from pragmatic reasoning processes. The human data revealed an interaction between garden-path and verb type, with higher error rates for OT verbs in non-garden-path conditions. This pattern suggests that when pragmatic reasoning is structurally permissible, humans rely on it considerably. Compared to these human results, LLMs appear to engage relatively less in pragmatic reasoning. This pattern held consistently across both yes/no (Lee and Shin 2025) and paraphrasing tasks.

A second difference relates to how task format affects performance. When examining human data, accuracy was lower in the paraphrasing task (Patson et al. 2009) than in the forced-choice paradigm (Christianson et al. 2001) for both garden-path and non-garden-path conditions, with greater declines observed in garden-path sentences. This decline is not unexpected because free recall may require more cognitive resources than forced-choice responses that provide explicit retrieval cues. In contrast to these human results, LLMs showed higher accuracy in the paraphrasing task, particularly GPT-4. In prior research, GPT-4 generally outperformed GPT-3.5 in non-garden-path conditions but failed to show this advantage in garden-path conditions. However, in the paraphrasing task, GPT-4 demonstrated better performance in garden-path conditions. For instance, GPT-4 showed 0% and 20% accuracy rates for OT verbs and RAT verbs respectively in the yes/no experiments (Lee and Shin 2025), but achieved 35% full analysis for both verb conditions in the paraphrasing task. The shift to the paraphrasing task appears to have enabled LLMs, particularly GPT-4, to process syntax more accurately.

On the surface, GPT-4's decreased error rates in the paraphrasing task appear paradoxical, as this task might be considered more challenging. However, whether the paraphrasing task also poses greater difficulty for LLMs remains unclear. Humans may find it more challenging due to the necessity for free recall, in contrast to forced-choice tasks that provide specific cues for memory retrieval (Cleary et al. 2018). LLMs might experience similar constraints as they are not completely immune to a type of 'memory loss.' In transformer architectures, as sentences get longer, the model must distribute its attention to each word across more positions in the sentence,

which can weaken the connections between distant elements (Vaswani et al. 2017). Another potential factor is the higher computational effort required for LLMs to produce full sentences rather than simple binary yes/no responses. However, given ChatGPT’s proficiency in managing extended dialogues (Achiam et al. 2023), such information loss or additional computational load from generating full sentences appears unlikely with the experiment’s brief sentences. If memory loss or computational load were relevant factors, error rates should increase in non-garden-path conditions as well. However, error rates in non-garden-path conditions actually decreased for both GPT-3.5 and GPT-4. Therefore, the paraphrasing task may not present the same difficulty for LLMs that it does for humans.

The paraphrasing task might actually be easier for LLMs from a processing perspective, which could explain GPT-4’s improved performance in garden-path conditions. Yes/no questions, despite being binary and seemingly straightforward, require generating a response based on new information not present in the given sentence. In contrast, paraphrasing involves modifying existing words, which could potentially be accomplished through simple synonym substitution rather than genuine comprehension and reformulation. This approach would be particularly effective in comma-separated non-garden-path conditions, where mere one-to-one synonym replacement could yield correct answers. In fact, both models showed improvement in the non-garden-path OT verb conditions, with error rates dropping from approximately 75% for GPT-3.5 and 50% for GPT-4 in forced-choice experiments to 11% and 2% respectively in the paraphrasing task. From this interpretation, the paraphrasing task may have been easier overall, leading to general performance improvements.

Nevertheless, synonym substitution alone cannot fully explain the performance patterns. If models relied on one-to-one word replacement, this strategy should succeed in garden-path conditions as well, since these sentences differ from non-garden-path conditions only in the absence of a comma and therefore permit the same synonym-based approach. Yet neither model showed improvements in garden-path conditions comparable to their gains in non-garden-path conditions. Thus, while some synonym substitution may have occurred, the overall approach likely involved genuine sentence reformulation. From this interpretation, GPT-4’s improved performance in garden-path conditions may suggest the presence of mechanisms that enable more accurate syntactic processing in this task format, rather than simply reflecting easier task demands.

One possible explanation for the difference in GPT-4’s performance between the two tasks may lie in how LLMs process sentences depending on task structure. In the prior yes/no question experiments, questions targeted specific parts of the sentence such as “hunted” and “deer.” For such tasks, it might be computationally efficient for LLMs to concentrate on the target words highlighted in the questions while paying less attention to other sentence parts. Given the attention mechanism utilized by LLMs, they likely process sentences centered around these specified target words. However, this processing strategy, while effective for answering yes/no questions, might heighten vulnerability to locally coherent but globally incorrect structures in garden-path sentences (*e.g.* “the man hunted the deer”). In contrast, the paraphrasing task demands processing of the entire sentence, leading LLMs to consider the full context rather than focusing solely on target words. Since the paraphrasing task necessitates a holistic approach to sentence processing, there is less incentive to disproportionately attend to specific parts of sentences. This processing method, which incorporates information from later sentence parts where disambiguating information appears, could lead to reduced susceptibility to garden-path errors.

Both GPT-3.5 and GPT-4 exhibited this task-dependent processing pattern. In the non-garden-path condition with OT verbs, where both models previously recorded high error rates in the yes/no questions task, they showed almost no errors in the paraphrasing task. However, GPT-3.5 continued to produce errors in garden-path conditions at rates comparable to those observed in yes/no questions, albeit with slight improvement. This

suggests that GPT-3.5 might possess weaker syntactic processing capabilities than GPT-4. Even when the later parts of sentences are taken into account in its processing, GPT-3.5 still struggled to overcome incorrectly formed relationships earlier in the sentence. On the other hand, when GPT-4 processes information from later sentence parts together, it demonstrates syntactic processing that is comparable to or in some aspects more accurate than human processing.

The task-dependent processing patterns observed in LLMs may be an inherent consequence of their fundamental processing mechanisms. LLMs are trained through next-token prediction based on input sequences, meaning they cannot generate outputs without corresponding inputs (Brown et al. 2020, Radford et al. 2019, Vaswani et al. 2017). This input-dependent architecture makes task-specific variation a natural feature rather an unexpected behavior. LLMs, particularly those deployed as conversational chatbots, are designed to interact with users, necessitating algorithms optimized for responding effectively to diverse task demands. From this perspective, variation in processing strategies according to task structure represents an expected outcome of their design.

Intriguingly, this task-dependent processing is not a unique characteristic of LLMs but is also observed in human language processing. Numerous studies have shown that human language processing can vary depending on the given task or type of question (Caplan et al. 2008, Caplan 2010, Franck et al. 2015, Gilbert et al. 2021, Salverda et al. 2011, Qian et al. 2018). For instance, Caplan et al. (2008) demonstrated that different brain regions are activated for syntactic processing depending on the type of task, and Qian et al. (2018) showed that changes in the type of question could alter understanding of garden-path sentences. These task effects highlight that human syntactic processing can be either deep or shallow, depending on the set goals. This aligns with the ‘good-enough’ approach of syntactic processing, which posits that humans do not always process syntax in a fully specified, detailed, complete, and accurate manner, but rather to the extent necessary for understanding, that is, to a ‘good-enough’ level. In this regard, task-dependent processing is not unique to LLMs but reflects a general feature of good-enough processing observed in both LLM and human language processing.

Nevertheless, LLMs differ from humans in ways that may render them more vulnerable to over-optimization. While humans also respond to external stimuli, task instructions, and conversational cues, they can additionally engage in spontaneous, proactive language generation based on autonomous observation and thought. In contrast, LLMs operate exclusively in a responsive mode, generating output only when prompted by input. This purely reactive nature, while generally efficient, may cause inefficiencies under certain conditions. Over-optimization in an attempt to produce rapid responses could paradoxically lead to suboptimal outcomes. By focusing too narrowly on generating immediate replies, LLMs may sometimes overlook broader contextual cues or deeper linguistic processing essential for more complex interactions, as suggested by the relatively higher error rates observed in the forced-choice question paradigm in prior research.

Despite these insights into LLM syntactic processing, the present study has several limitations. First, we examined only specific models from OpenAI’s GPT family, limiting generalizability across different LLM architectures and newer model versions. Second, the opacity of these models’ internal processes prevented direct examination of attention mechanisms or internal representations that might reveal underlying syntactic processing mechanisms. Third, while we employed a single prompt formulation to maintain consistency with the human experimental paradigm from Patson et al. (2009), different paraphrasing instructions might yield different results.

Future research should address these limitations by testing diverse model architectures, exploring various prompt formulations, and employing methods to probe internal model representations. Additionally, investigating how different task designs systematically affect processing depth could inform the development of more robust evaluation frameworks and more effective prompt engineering strategies.

## 5. Conclusion

This study demonstrates that good-enough processing in LLMs represents genuine syntactic processing characteristics rather than methodological artifacts. LLMs exhibit human-like incomplete reanalysis patterns across both forced-choice and paraphrasing paradigms, despite their substantial computational resources. Moreover, we also found significant task-dependent variation in processing depth. GPT-4 showed improved performance in paraphrasing compared to forced-choice tasks, suggesting that advanced capabilities manifest differently across task structures. These patterns align with human good-enough processing, where comprehenders settle for interpretations that are adequate rather than completely accurate.

These findings carry practical implications for LLM development and application. Current evaluation benchmarks may not capture the full range of model capabilities, as performance varies across task formats even when testing the same underlying construct. This suggests a need for diverse evaluation paradigms that assess syntactic processing through multiple task structures rather than relying on single formats. From a development perspective, models may benefit from training procedures that explicitly address multiple task formats, helping to balance efficiency with processing depth while mitigating risks of over-optimization for immediate responses. For users and practitioners, these results highlight the role of prompt engineering in eliciting different levels of processing. The way tasks are structured and questions are framed can substantially alter the depth of syntactic analysis that models perform. This means that response quality depends not only on model capability but also on task design choices. Even advanced models demonstrate task-dependent variation in processing depth, suggesting that applications requiring reliable syntactic understanding may benefit from employing multiple query formats to verify comprehension. Understanding these task-dependent processing patterns is essential for effective deployment of LLMs in contexts where syntactic accuracy matters.

## References

- Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, ... and B. McGrew. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Ardoin, T., Y. Cai and G. Wunder. 2025. Where confabulation lives: Latent feature discovery in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 29801-29825.
- Bacon, G. and T. Regier. 2019. Does BERT agree? Evaluating knowledge of structure dependence through agreement relations. *arXiv preprint arXiv:1908.09892*.
- Bender, E. M., T. Gebru, A. McMillan-Major and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.
- Bernardy, J. P. and S. Lappin. 2017. Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology* 15(2), 1-15.
- Bever, T. G. 1970. The cognitive basis for linguistic structures. In J. R. Hayes, ed., *Cognition and the Development of Language*, 279-362. New York: Wiley and Sons.
- Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, ... and D. Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33, 1877-1901.

- Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, ... and Y. Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint* arXiv:2303.12712.
- Caplan, D. 2010. Task effects on BOLD signal correlates of implicit syntactic processing. *Language and Cognitive Processes* 25(6), 866-901.
- Caplan, D., E. Chen and G. Waters. 2008. Task-dependent and task-independent neurovascular responses to syntactic processing. *Cortex* 44(3), 257-275.
- Chaves, R. P. 2020. What don't RNN language models learn about filler-gap dependencies? In *Proceedings of the Society for Computation in Linguistics* 3(1), 20-30.
- Chowdhury, S. A. and R. Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, 133-144.
- Christianson, K., A. Hollingworth, J. Halliwell and F. Ferreira. 2001. Thematic roles assigned along the garden path linger. *Cognitive Psychology* 42(4), 368-407.
- Cleary, A. M., A. J. Ryals and J. S. Nomi. 2018. Dependent measures in memory research: From free recall to recognition. In *Handbook of Research Methods in Human Memory*, 19-35. New York: Routledge.
- Davis, E. and G. Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM* 58(9), 92-103.
- Dreyfus, H. L. 1972. *What Computers Can't Do: The Limits of Artificial Intelligence*. New York: Harper and Row.
- Farquhar, S., J. Kossen, L. Kuhn and Y. Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 625-630.
- Ferreira, F., K. G. D. Bailey and V. Ferraro. 2002. Good-enough representations in language comprehension. *Current Directions in Psychological Science* 11(1), 11-15.
- Ferreira, F. and C. Clifton, Jr. 1986. The independence of syntactic processing. *Journal of Memory and Language* 25(3), 348-368.
- Ferreira, F. and N. D. Patson. 2007. The 'good enough' approach to language comprehension. *Language and Linguistics Compass* 1(1-2), 71-83.
- Frances, C. 2024. Good-enough language processing: A satisficing approach to language comprehension and production. *Language and Linguistics Compass* 18(1), e12513.
- Franck, J., S. Colonna and L. Rizzi. 2015. Task-dependency and structure-dependency in number interference effects in sentence comprehension. *Frontiers in Psychology* 6, 349.
- Frazier, L. and K. Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14(2), 178-210.
- Futrell, R., E. Wilcox, T. Morita and R. Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint* arXiv:1809.01329.
- Gibson, E. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68(1), 1-76.
- Gibson, E. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz and W. O'Neil, eds., *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, 95-126. Cambridge, MA: MIT Press.
- Gilbert, R. A., M. H. Davis, M. G. Gaskell and J. M. Rodd. 2021. The relationship between sentence comprehension and lexical-semantic retuning. *Journal of Memory and Language* 116, 104188.
- Goldberg, Y. 2019. Assessing BERT's syntactic abilities. *arXiv preprint* arXiv:1901.05287.
- Gulordava, K., P. Bojanowski, É. Grave, T. Linzen and M. Baroni. 2018. Colorless green recurrent networks

- dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195-1205.
- Hauser, M. D., N. Chomsky and W. T. Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298(5598), 1569-1579.
- Hu, J., J. Gauthier, P. Qian, E. Wilcox and R. Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1725-1744.
- Lampinen, A. K., I. Dasgupta, S. C. Chan, H. R. Sheahan, A. Creswell, D. Kumaran, ... and F. Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus* 3(7), pgae233.
- Lee, J. and J.-A. 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, J. and J.-A. Shin. 2025. Good-enough but more error-prone: Garden-path processing in GPT models. *Linguistic Research* 42(3), 539-579.
- Lee, J., J.-A. Shin and M. K. 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.
- Levy, R. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3), 1126-1177.
- Linzen, T. and N. Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 690-695.
- Linzen, T., E. Dupoux and Y. Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4, 521-535.
- Loftus, E. F. and J. E. Pickrell. 1995. The formation of false memories. *Psychiatric Annals* 25(12), 720-725.
- Madaan, L., D. Esiobu, P. Stenetorp, B. Plank and D. Hupkes. 2025. Lost in inference: Rediscovering the role of natural language inference for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 9229-9242.
- MacDonald, M. C. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology* 4, 226.
- MacDonald, M. C., M. A. Just and P. A. Carpenter. 1992. Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology* 24(1), 56-98.
- MacDonald, M. C., N. J. Pearlmutter and M. S. Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101(4), 676-703.
- Marvin, R. and T. Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1192-1202.
- Osterhout, L. and P. J. Holcomb. 1992. Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language* 31(6), 785-806.
- Patson, N. D., E. S. Darowski, N. Moon and F. Ferreira. 2009. Lingering misinterpretations in garden-path sentences: Evidence from a paraphrasing task. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35(1), 280-285.
- Qian, Z., S. Garnsey and K. Christianson. 2018. A comparison of online and offline measures of good-enough processing in garden-path sentences. *Language, Cognition and Neuroscience* 33(2), 227-254.
- Radford, A., J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1(8), 9.

- Rayner, K. and L. Frazier. 1987. Parsing temporarily ambiguous complements. *Quarterly Journal of Experimental Psychology* 39A(4), 657-673.
- Salverda, A. P., M. Brown and M. K. Tanenhaus. 2011. A goal-based perspective on eye movements in visual world studies. *Acta Psychologica* 137(2), 172-180.
- Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3), 417-424.
- Sharma, M., M. Tong, T. Korbak, D. Duvenaud, A. Aspell, S. Bowman, E. Durmus, Z. Hatfield-Dodds, S. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang and E. Perez. 2024. Towards Understanding Sycophancy in Language Models. In *Proceedings of International Conference on Representation Learning 2024*, 110-144.
- Shultz, T. R., J. M. Wise and A. S. Nobandegani. 2025. Text understanding in GPT-4 versus humans. *Royal Society Open Science* 12(2), 241313.
- Slattery, T. J., P. Sturt, K. Christianson, M. Yoshida and F. Ferreira. 2013. Lingerin misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language* 69(2), 104-120.
- Tian, Y., T. Huang, M. Liu, D. Jiang, A. Spangher, M. Chen, ... and N. Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17659-17681.
- Trueswell, J. C., M. K. Tanenhaus and S. M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language* 33(3), 285-318.
- Turpin, M., J. Michael, E. Perez and S. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36, 74952-74965.
- van Schijndel, M. and T. Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, 2603-2608.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30.
- Wilcox, E., R. Levy and R. Futrell. 2019. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 181-190.
- Wilcox, E., R. Levy, T. Morita and R. Futrell. 2018. What do RNN language models learn about filler-gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 211-221.

Examples in: English

Applicable Languages: English

Applicable Level: Tertiary