International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Comparative Study of ChatGPT and DeepSeek on Syntactic Ambiguity

Tingting Geng
College of Foreign Studies, Jinan University, Guangzhou, China
Corresponding Author: Tingting Geng, **E-mail**: gtt624725@163.com

| ABSTRACT
Linguistic ambiguity poses significant challenges to natural language processing (NLP) systems, particularly in context-dependent tasks such as machine translation and dialogue generation. While large language models (LLMs) like ChatGPT and DeepSeek exhibit advanced language capabilities, their efficacy in processing ambiguous structures remains understudied. This study constructs a systematically annotated corpus comprising 100 English sentences and two supplementary test sets, focusing primarily on three types of syntactic ambiguity: prepositional phrase (PP) attachment, attributive modification, and coordination scope. Adopting a dual-phase experimental design, we evaluate and compare the capabilities of ChatGPT-5 Mini and DeepSeek-R1 across two dimensions: (1) ambiguity detection accuracy tested through 100 separate dialogues to ensure controlled experimental conditions, and (2) disambiguation performance assessed by comparing model outputs across different contextual conditions. The results indicate that: (1) both models demonstrate strong competence in ambiguity detection, with DeepSeek-R1 achieving an accuracy of 85%, marginally outperforming ChatGPT-5 Mini (77%); (2) in context-dependent disambiguation tasks, neither model effectively utilizes contextual information to resolve PP attachment ambiguities; (3) similarly, both models fail to leverage contextual cues to disambiguate attributive modification structures. These findings suggest that LLM performance may rely more on surface-level statistical patterns than on deep syntactic reasoning, particularly struggling with non-local dependency resolution and probabilistic attachment interpretation. Furthermore, this study contributes to LLM interpretability research and provides implications for developing ambiguity-aware NLP applications.

| KEYWORDS

ChatGPT; DeepSeek; syntactic ambiguity; ambiguity detection; disambiguation performance

| ARTICLE INFORMATION

ACCEPTED: 01 July 2026 **PUBLISHED:** 15 August 2026 **DOI:** 10.32996/ijllt.2026.9.8.16

1. Introduction

Since the "linguistic turn" of the 20th century, the philosophy of language has been centrally concerned with the relationship between language, meaning, and understanding. Within this tradition, different theoretical frameworks have offered competing accounts of how linguistic meaning is represented and processed. For instance, Wittgenstein (2009) emphasizes the role of use and context in meaning construction, whereas Chomsky (2002) proposes that language is grounded in an innate system of abstract grammatical rules. These contrasting perspectives continue to inform contemporary debates on the nature of language comprehension.

Recent advances in large language models (LLMs), such as ChatGPT and DeepSeek, have renewed interest in these longstanding issues. Although such models demonstrate high levels of fluency and coherence across a wide range of natural language tasks, it remains unclear whether their performance reflects genuine language understanding or the exploitation of statistical regularities in large-scale data. This question has become a central concern in both computational linguistics and the philosophy of language.

Syntactic ambiguity provides a useful empirical domain for investigating this issue. As Quine (2013) argues, linguistic expressions often admit multiple interpretations, highlighting the indeterminacy inherent in language. For example, sentences involving prepositional phrase attachment (e.g., "I saw the man with the telescope") can be interpreted in structurally distinct ways. Human language users typically resolve such ambiguities by integrating syntactic, semantic, and contextual information. This integrative process is widely regarded as a key indicator of systematic and productive language understanding. In contrast, it remains an open question whether LLMs can represent and resolve such ambiguities in a comparable manner.

Existing research has begun to evaluate the linguistic capabilities of LLMs from a cognitive and computational perspective. For example, Cai et al. (2024) report that ChatGPT exhibits human-like performance in tasks such as structural priming and implicit causal reasoning, but shows clear limitations in using contextual information to resolve syntactic ambiguities. Similarly, Chomsky et al. (2023) argue that LLMs primarily rely on statistical pattern matching rather than rule-governed syntactic computation. In contrast, emergentist accounts (e.g., Piantadosi, 2023) suggest that complex linguistic knowledge may arise from large-scale data and model capacity. Despite these differing interpretations, there is broad agreement that the ability of LLMs to handle ambiguity—particularly in context-dependent settings—remains insufficiently understood.

Against the backdrop, the present study conducts a systematic comparative investigation of syntactic ambiguity processing in LLMs, focusing on ChatGPT-5 Mini and DeepSeek-R1. Adopting an experimental approach, the study examines two core aspects of model performance: (1) the accuracy of ambiguity detection, and (2) the ability to resolve ambiguity under varying contextual conditions. Specifically, the study addresses the following research questions: (1) How accurately can LLMs detect syntactic ambiguity? (2) To what extent can they utilize contextual information to resolve prepositional phrase attachment and attributive modification ambiguities?

2. Literature Review

2.1 Philosophical Perspectives on Language Understanding

The nature of language understanding has been a central concern in both philosophy and linguistics, providing an important theoretical foundation for evaluating artificial language systems. Different frameworks offer contrasting accounts of how meaning is constructed and processed. From a usage-based perspective, Wittgenstein (2009) argues that meaning is not an inherent property of linguistic forms but is constituted through use within specific social and contextual practices. This view highlights the importance of context, interaction, and pragmatic inference in language comprehension. In contrast, Chomsky's generative grammar framework (2002, 2014) posits that language is grounded in an innate system of abstract syntactic rules, often referred to as universal grammar. Within this framework, language understanding is primarily a function of internal grammatical competence.

Unlike the above two, Quine (2013) introduces a different line of argument by emphasizing the indeterminacy of reference and translation. He suggests that linguistic meaning is fundamentally underdetermined by observable data, implying that multiple interpretations may always be compatible with the same linguistic input. This perspective underscores the inherent ambiguity of language and raises challenges for any system that attempts to model understanding. More recently, connectionist approaches represented by Hinton (2023) propose that language learning can emerge from large-scale statistical learning without explicitly encoded grammatical rules. In this framework, linguistic knowledge is represented as distributed patterns in high-dimensional vector spaces, and learning is achieved through exposure to large amounts of data.

Taken together, the aforementioned philosophical divisions have established a rich theoretical background for examining large language models: When evaluating these systems, are they more akin to the "language game participants" described by Wittgenstein or do they potentially embody the generative grasp of abstract grammatical rules emphasized by Chomsky? Or, do their "black box" characteristics and the frequent inconsistencies in their outputs fundamentally confirm Quine's profound insight into the uncertainty of meaning and understanding? And the connectionist program of Hinton et al. also forces us to consider whether there might exist a third form of "understanding"—that is, intelligent intelligence emerging from massive data based on distributed representations. This series of inquiries provide a solid philosophical motivation and analytical framework for this study to empirically test the understanding mechanism of large language models by starting from specific syntactic ambiguity phenomena.

2.2 Syntactic Ambiguity and Language Processing

Syntactic ambiguity has long been a focal topic in both theoretical linguistics and psycholinguistics. It arises when a single surface structure admits multiple underlying syntactic interpretations, as in the case of prepositional phrase (PP) attachment or coordination scope ambiguity (Frazier & Rayner, 1982). From a cognitive perspective, ambiguity resolution is a key aspect of real-time language comprehension. Early models such as the garden-path theory (Frazier, 1978) propose that sentence

processing relies on heuristic strategies like minimal attachment. In contrast, constraint-based approaches (e.g., Tanenhaus et al., 1995) argue that multiple sources of information—including lexical semantics, discourse context, and world knowledge—are integrated dynamically during processing.This distinction is particularly relevant for evaluating artificial systems. The ability to resolve syntactic ambiguity requires not only structural representation but also the integration of contextual and semantic information. As such, ambiguity resolution serves as a sensitive diagnostic for assessing whether a system exhibits shallow pattern recognition or deeper integrative processing.

To operationalize this evaluation, computational research has introduced controlled benchmarks such as the Winograd Schema Challenge (Levesque et al., 2012) This paradigm, widely regarded as the "new Turing test" for semantic understanding (Marcus, 2017), embeds deep common sense reasoning within the task of pronoun resolution by designing "sentence pairs" that differ by only one or two "trigger words", aiming to render models relying on shallow statistical correlations ineffective. However, with the emergence of large language models, the distinction of traditional "sentence pairs" has faced challenges. To address this, Yuan (2024) proposed an important methodological improvement: by simultaneously replacing the "anchor words" and "trigger words" of sentences, constructing "sentence pairs" consisting of two sentences with more complex semantic relationships. For example, in the "保姆抱不动/一把抱起女孩，因为她太壮实/瘦弱" pair of sentences, the model can no longer rely on the co-occurrence statistics of a single word, but must understand the complex logical relationship between the action of "抱" and the attributes of "壮实/瘦弱". Cai et al. (2024) also adopted a similar research method in their study. They changed the original context and the probe type of the question, for instance, in the sentence pair "There was a hunter and a poacher/two poachers. The hunter killed the dangerous poacher with a rifle not long after sunset", when under different contexts (a poacher/two poachers), different models gave different answers to the questions like "Did the hunter use a rifle" and "Did the poacher use a rifle". These "sentence pair" paradigms significantly increase the semantic confusion degree, providing a more rigorous tool for detecting whether models perform true integrated reasoning, and also offering methodological inspiration for us to construct experimental materials that can effectively test whether models rely on superficial correlations or deeper interpretive mechanisms.

2.3 Evaluation of the Language Capabilities of LLMs

Recent studies have investigated the extent to which large language models exhibit human-like language processing behavior. For instance, Cai et al. (2024) report that models such as ChatGPT demonstrate similarities to human performance in tasks including structural priming and causal reasoning. However, they also identify systematic limitations in the use of contextual information for resolving syntactic ambiguities. Related work by Yuan (2024, 2025) provides further evidence that LLMs perform well on certain structured tasks, such as pronoun resolution and negation scope interpretation, but show instability when deeper semantic integration is required. In particular, performance tends to degrade in tasks involving complex compositional reasoning or reliance on world knowledge. These findings suggest that LLMs exhibit an uneven profile of linguistic competence: while they can capture statistical regularities at the surface level, their ability to integrate multiple sources of information in a coherent and context-sensitive manner remains limited. This limitation is especially evident in tasks that require resolving ambiguity across sentence and discourse levels.

Despite these advances, the existing literature has not yet provided a systematic comparative evaluation of different types of syntactic ambiguity within a unified experimental framework. In particular, the interaction between ambiguity detection and context-dependent disambiguation remains underexplored. Building on the theoretical and empirical work reviewed above, the present study aims to address this gap by conducting a controlled comparative analysis of syntactic ambiguity processing in large language models. By focusing on both ambiguity detection and context-sensitive resolution across multiple ambiguity types, this study seeks to provide a more fine-grained account of the strengths and limitations of LLMs in handling structurally complex language.

3. Methodology

3.1 Construction of the Corpus

This study adopts a controlled experimental design to investigate how large language models (LLMs) process syntactic ambiguity. A comparative approach is used to evaluate two models—ChatGPT-5 Mini and DeepSeek-R1—across two core tasks: (1) ambiguity detection and (2) ambiguity resolution. The design aims to assess both the models' ability to recognize structurally ambiguous sentences and their capacity to resolve such ambiguity using contextual information. So to ensure the systematic and targeted nature of the assessment, this study constructed three specialized corpora with different functional orientations, corresponding to specific research questions. The selection and construction of all the corpora were guided by linguistic theories and aimed to strike a balance between experimental control and ecological validity.

Firstly, to assess the model's ability to recognize syntactic ambiguity (research question one), a diagnostic corpus consisting of 100 English sentences was constructed. Among them, 80 were test sentences containing various types of syntactic ambiguity, and 20 were control sentences with clear syntactic structures and no ambiguity. The types of ambiguity in the test sentences covered the three classic categories in linguistic research: prepositional phrase, attachment ambiguity, attributive modification ambiguity, and coordinate structure scope ambiguity (Carnie, 2021). The corpus sources included examples from classic linguistic textbooks and typical sentences extracted from relevant computational linguistics and psycholinguistics research papers.

Secondly, to deeply explore the disambiguation performance of the model in specific contexts (research question two), two refined test sets were also constructed . The first one is the prepositional phrase attachment ambiguity test set, which directly adopted 32 core materials that had been verified by Cai et al. (2023) in their psycholinguistic experiment. Each group of materials was derived by systematically manipulating two variables, "context" (single referent vs. multiple referents) and "probe type" (VP probe vs. NP probe), resulting in 4 test sentences (see Example 1), totaling 128 sentences. This design can effectively test whether the model, like humans, uses the referential information in the text to parse the attachment relationship. The second is the attributive modification ambiguity test set. We strictly followed the experimental paradigm and logic of Cai et al. (2023) and newly constructed 32 parallel materials. By replacing core nouns, relative clause verbs, and adjectives, we generated 128 test sentences with structurally equivalent but completely new lexical contents. For Example 2, in the prototype "The teacher praised the student of the professor who was enthusiastic", the key components are systematically replaced to test the model's general processing ability for the abstract syntactic relationship of distant relative clause modification, rather than its memory of specific lexical combinations.

Example 1.
Single Referent, NP Probe: There was a knight and a king. The knight struck the king with a sword during the dark night. Did the knight use a sword, yes or no?
Single Referent, VP Probe: There was a knight and a king. The knight struck the king with a sword during the dark night. Did the struck king have a sword, yes or no?
Multiple Referents, NP A Probe: There was a knight and two kings. The knight struck the king with a sword during the dark night. Did the knight use a sword, yes or no?
Multiple Referents, NP B Probe: There was a knight and two kings. The knight struck the king with a sword during the dark night. Did the struck king have a sword, yes or no?

Example 2.
Single Referent, NP A Probe: There was a student and a professor. The teacher praised the student of the professor who was enthusiastic. Was the student enthusiastic, yes or no?
Single Referent, NP B Probe: There was a student and a professor. The teacher praised the student of the professor who was enthusiastic. Was the professor enthusiastic, yes or no?
Multiple Referents, NP A Probe: There were two students and a professor. The teacher praised the student of the professor who was enthusiastic. Was the student enthusiastic, yes or no?
Multiple Referents, NP B Probe: There were two students and a professor. The teacher praised the student of the professor who was enthusiastic. Was the professor enthusiastic, yes or no?

3.2 Experimental Design

This study employed a two-stage experimental design to successively examine the model's ability to identify static ambiguities and to disambiguate in dynamic contexts.

The first stage is ambiguity detection. In this phase, the aforementioned 100 sentence corpus is used to test the model through a binary classification task. To ensure the independence of each query and prevent the model from obtaining clues from the conversation history, we conduct tests on each sentence in the corpus by initiating a new conversation session. The prompt uniformly requires the model to determine the target sentence:

Does the following sentence have more than one possible grammatical or meaning interpretation? Please answer only with "Yes" or "No".
Sentence: [...]

So 100 independent tests were conducted to test on both ChatGPT-5 Mini and DeepSeek-R1 to obtain the baseline accuracy rate for their ambiguity detection.

The second stage is the evaluation of disambiguation performance. In this phase, two disambiguation test sets (each consisting of 128 sentences) are used. The experiment adopts the context-based forced selection paradigm. Each test item begins with a short context paragraph, which clearly specifies the number of referents (single or multiple), followed by the target sentence with ambiguity, and finally a probing question pointing to a specific syntactic interpretation (see Example 3) . The model needs to conduct reasoning based on the entire discourse paragraph and provide a "yes" or "no" answer. By systematically comparing the answer patterns of the model under different experimental conditions (such as NP detection in multi-referential contexts vs. NP detection in single-referential contexts), we can precisely quantify the extent to which contextual information affects the model's disambiguation decisions, thereby determining whether its processing process relies on deep syntactic-discourse integration or is constrained by superficial lexical associations or problem phrasing deviations.

Example 3.
There was a student and a professor. The teacher praised the student of the professor who was enthusiastic. Was the student enthusiastic, yes or no?

3.3 Model and Implementation

This study selected ChatGPT-5 Mini (OpenAI) and DeepSeek-R1 (DeepSeek) as the comparison objects. Both are current leading large language models with representative architectures. All experiments were completed through the official API interface in November 2025. To ensure the reliability of the results, standardized and neutral prompt templates were designed to avoid introducing extraneous guidance. All model outputs were fully recorded, and subsequent quantitative statistical analysis (such as accuracy rate, logistic regression analysis) and qualitative in-depth analysis of key cases will be conducted to comprehensively reveal their capability advantages and inherent limitations.

4. Results & Discussion

4.1 Accuracy of Ambiguity Detection

The first stage of the study examined the ability of ChatGPT-5 Mini and DeepSeek-R1 to detect syntactic ambiguity. Using a dataset of 100 English sentences (80 ambiguous and 20 unambiguous controls), both models performed a binary classification task to determine whether each sentence admitted more than one interpretation.The results showed that both models significantly outperformed the random guessing level (50%), proving that large language models can acquire certain patterns related to syntactic ambiguity from the training data. Specifically, DeepSeek-R1 achieved a comprehensive accuracy rate of 85%, while ChatGPT-5 Mini's accuracy rate was 77%, indicating that the former has a certain advantage in this task (see Table 1).

Table 1. Detection Accuracy Results

Model	Overall Accuracy (N=100)
DeepSeek-R1	85%
ChatGPT-5 Mini	77%

Despite this above-chance performance, a gap remains when compared to human benchmarks reported in previous studies, where accuracy typically approaches or exceeds 95% (Cai et al., 2023; Yuan, 2024) (see Figure 1). This difference suggests that, although large language models are capable of identifying common ambiguity patterns, their performance does not fully align with human linguistic judgments. Also, qualitative analysis of model errors further reveals systematic tendencies rather than random misclassification. In particular, both models tend to classify less prototypical or context-dependent ambiguities as unambiguous. This pattern suggests that model judgments rely more heavily on frequent or salient surface cues than on abstract structural representations that generalize across contexts. When ambiguity cues are less common or less explicitly marked, detection accuracy declines.

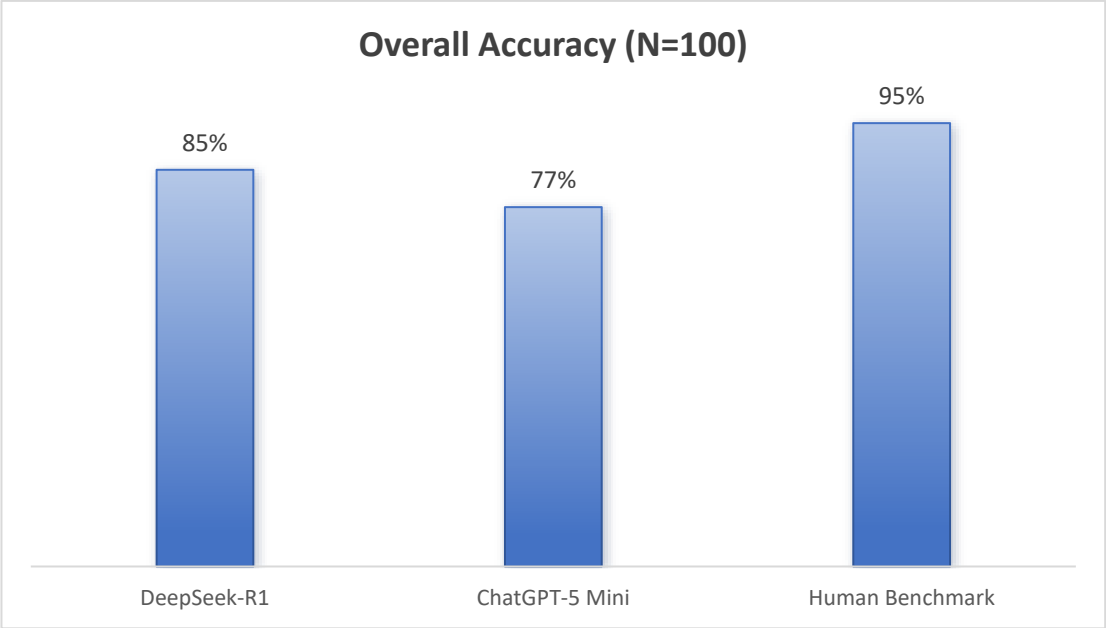

Figure 1. Detection Accuracy Results

The above findings also have significant theoretical implications. First, it partially validates the criticism made by Chomsky et al. (2023) regarding LLMs as "advanced pattern matchers" rather than "understanders". The fact that the model performs moderately well but not exceptionally in ambiguity detection indicates that its success may stem from its massive memory and imitation of the "ambiguity-surface cues" associations in vast amounts of text, rather than mastering abstract syntactic rules that can recursively generate and recognize all possible structures. At the same time, this also echoes the core concern raised by Liu et al. (2023) in their research: LLMs tend to output a seemingly certain and fluent interpretation, often ignoring or suppressing the inherent uncertainty of language. Our experiments at the syntactic level confirm this tendency towards "insensitivity to ambiguity" or "overdetermination".

In conclusion, the ambiguity detection task serves as a preliminary diagnostic tool, effectively revealing a facet of the language capabilities of LLMs: they possess a statistically-based, quite powerful pattern recognition ability, capable of handling many common language phenomena. However, their underlying mechanism differs qualitatively from human judgments based on rules and understanding. This provides a crucial preparatory cognitive and logical starting point for further in-depth analysis of their performance in more dynamic tasks that require active disambiguation.

4.2 Resolution of PP Attachment Ambiguity

In the disambiguation test for the PP attachment ambiguity, neither of the two models demonstrated the human-like ability to flexibly utilize contextual information. However, the two models exhibit distinct response patterns.

4.2.1 ChatGPT-5 Mini: Reliance on Surface-level Clues

In the task of PP attachment disambiguation, the performance of ChatGPT-5 Mini exhibits a mechanical processing mode that is highly dependent on surface linguistic cues and almost disregards deep discourse information. Statistical analysis provides clear quantitative evidence for its decision-making mechanism. Logistic regression analysis shows (see Table 2 and Table 3), the effect of contextual factors (single referent / multiple referents) is not significant ($p = 0.340$), indicating that the model fails to effectively utilize the referential information in the discourse to regulate its syntactic analysis. In sharp contrast, the type of the probing question (VP probe / NP probe) has a significant statistical effect ($p = 0.042$, OR = 5.127), which quantitatively confirms that the surface wording of the question is the most dominant factor driving the model's response. Further observation of the proportion distribution of its NP attachment judgment reveals that when the probing question directly asks about the noun phrase (NP Probe), the proportion of the model's choice for NP attachment (12.500% to 15.625%) is significantly higher than when the question asks about the verb phrase (VP Probe) (0% to 6.250%). This pattern indicates that the processing of ChatGPT is more like a "keyword triggering" mechanism: it isolates and analyzes the core verb (such as "use") or noun in the question and makes shallow associations with the corresponding components in the sentence, rather than evaluating the relative rationality of different syntactic interpretations based on an integrated understanding of the entire discourse context.

Table 2. Binary Logistic Regression Analysis for ChatGPT-5Mini on PP Attachment Ambiguity Task

Predictor	B(Coefficient)	SE	Wald	*df*	P-value	OR
Context	0.635	0.666	0.909	1	0.340	1.887
Probe Type	1.634	0.806	4.117	1	0.042*	5.127
Constant	-3.798	0.838	20.550	1	0.001*	0.022

Note: The dependent variable is attachment type and ***, **, * indicate significance levels of 1%, 5%, and 10% respectively.

Table 3. Proportion of NP Attachment Judgments by ChatGPT-5Mini

Probe Type	Context	n	Proportion of NP Attachment
VP Probe	Single Referent	32	0.000%
	Multiple Referents	32	6.250%
NP Probe	Single Referent	32	12.500%
	Multiple Referents	32	15.625%

This disregard for contextual cues is particularly prominent in specific cases. As shown in Examples 4 and 5, while maintaining the same context, changing the probe type, ChatGPT consistently responds with "yes". This answer pattern that ignores logical reasoning and discourse constraints profoundly reveals the systematic flaws in the current large language model's integration at the syntactic-semantic-pragmatic interface. It suggests that the model's operation is not based on constructing a dynamic and coherent mental model, but rather is more likely to be searching within its vast parameter space for the most frequent statistical correlations related to specific word combinations (such as the "with" phrase and "use"). This finding resonates with the criticisms from the theoretical community, namely that such models are essentially "cumbersome statistical engines" (Chomsky et al., 2023), and behind their smooth outputs, there is a lack of the core cognitive ability of humans to flexibly weigh multiple information sources during the process of disambiguation. At the same time, it also confirms the concern that large language models tend to suppress the inherent uncertainty of language and output an interpretation that seems certain but may violate contextual logic (Liu et al., 2023). Therefore, ChatGPT's failure in the PP attachment disambiguation task is not a random error, but a necessary manifestation of its fundamental limitation of relying on surface statistical patterns rather than deep integrated reasoning.

Example 4.
Single referent, VP probe: There was an explorer and a tent. The explorer found the tent with charts after a long search. Did the explorer use charts, yes or no? [yes]
Single referent, NP probe: There was an explorer and a tent. The explorer found the tent with charts after a long search. Did the tent use charts, yes or no? [yes]

Example 5.
Single referent, VP probe: There was a hooligan and two shops. The hooligan damaged the shop with fireworks late at night. Did the hooligan use fireworks, yes or no? [yes]
Single referent, NP probe: There was a hooligan and two shops. The hooligan damaged the shop with fireworks late at night. Did the damaged shop have fireworks, yes or no? [yes]

4.2.2 DeepSeek-R1: Limited Sensitivity and Non-integration Strategies

In contrast, DeepSeek-R1 shows a more complex pattern of behavior. Quantitative analysis reveals the dual influence pattern of its decision-making mechanism. Logistic regression analysis indicates (see Table 4), both the context and the type of the detection question have statistically significant effects on its decision-making (context: p = 0.014, OR = 4.100; question type: p = 0.002, OR = 8.086). This result is completely different from the situation where the contextual effect is not significant in ChatGPT, indicating that DeepSeek can perceive and respond to changes in referential information at the discourse level. However, further analysis of attachment proportions indicates that the model adjusts its responses in line with contextual variation to some extent. For example, in a single referential context, the proportion of NP attachment judgments for NP detection questions by the model is 15.625%, while in a multi-referential context, this proportion significantly increases to 37.500% (see Table 5). This trend of change with the context, especially the significant increase in the probability of NP attachment in multi-referential contexts, indicates that the model does not completely ignore the context.

Table 4. Binary Logistic Regression Analysis for DeepSeek-R1 on PP Attachment Ambiguity Task

Predictor	**B(Coefficient)**	**SE**	**Wald**	***df***	**P-value**	**OR**
Context	1.414	0.577	5.983	1	0.014**	4.100
Probe Type	2.090	0.669	9.755	1	0.002***	8.086
Constant	-3.931	0.749	27.545	1	0.000***	0.020

Table 5. Proportion of NP Attachment Judgments by DeepSeek-R1

Probe Type	**Context**	**n**	**Proportion of NP Attachment**
VP Probe	Single Referent	32	0.000%
	Multiple Referents	32	9.375%
NP Probe	Single Referent	32	15.625%
	Multiple Referents	32	37.500%

However, this sensitivity remains limited. Firstly, the influence of the question type (OR = 8.086) is still much greater than that of the context (OR = 4.100), indicating that the dominant factor in the model's decision is still the surface wording of the question. Secondly, more crucially, the model does not seem to integrate the contextual information and syntactic interpretation in a logically coherent manner as humans would. For Example 6: in a multi-referent context, when the probing question is NP probe, human readers naturally interpret "with a heater" as modifying "the room" to distinguish the target room from another room, thus being more inclined to answer "yes". In this condition, the NP attachment proportion (37.5%) of DeepSeek-R1 has increased, but it is far from reaching the high consistency that humans might exhibit. This indicates that the model may not truly understand the logic of the discourse-syntactic interface that "multi-reference contexts require more precise referential identification, thereby enhancing the rationality of PP modifying NP". Instead, its behavioral pattern is more in line with a "simplified addition" strategy: The model may separately calculate the weight of the "multi-reference" context cues and the weight of the "NP probing" question cues, and then mechanically add the two together, rather than like humans, performing nonlinear reasoning based on constraint satisfaction in a unified discourse representation and dynamically choosing the most coherent and reasonable interpretation (Linzen & Baroni, 2021).

Example 6.

Single referent, VP probe: There was a landlady and a room. The landlady warmed the room with a heater within a couple of minutes. Did the landlady use a heater, yes or no? [yes]

Single referent, NP probe: There was a landlady and a room. The landlady warmed the room with a heater within a couple of minutes. Did the warmed room have a heater, yes or no? [no]
Multiple referents, VP probe: There was a landlady and two rooms.The landlady warmed the room with a heater within a couple of minutes. Did the landlady use a heater, yes or no? [yes]
Multiple referents, NP probe: There was a landlady and two rooms.The landlady warmed the room with a heater within a couple of minutes. Did the warmed room have a heater, yes or no? [yes]

Therefore, although the performance of DeepSeek-R1 is more complex than that of ChatGPT-5 Mini, demonstrating an initial ability to capture contextual cues, its underlying processing mechanism still has a fundamental gap from the elegant, flexible and highly integrated disambiguation reasoning of humans. It fails to integrate the discourse model, syntactic analysis and semantic rationality evaluation organically to determine "which interpretation is more reasonable in the given context". This gap between "sensitivity" and "integration" further confirms the systematic challenges faced by current large language models in core cognitive tasks that require dynamic coordination of multiple sources of information (Cai et al., 2024), and also suggests that purely scale expansion and the training goal of predicting the next word may not be sufficient to spontaneously emerge a similar human-like dynamic interface mechanism for syntax and discourse.

4.3 Resolution of Attribute Modification Ambiguity

As for the task of resolving ambiguity in attributive modifiers, the model exhibits another typical non-human pattern. Similar to the PP attachment task, the detection of problem types once again becomes the dominant factor, but the influence of the context is modulated in a more distorted manner.

4.3.1 Distorted Interaction Patterns

Both ChatGPT-5 Mini and DeepSeek-R1 failed to demonstrate the flexible and collaborative information integration ability of humans. Instead, they exhibited a distorted and non-linear interaction pattern. Data analysis shows that probe type and context jointly influence the model's decision-making, but the interaction between the two is not as synergistically promoting the most reasonable interpretation as humans understand it. Instead, it presents a distorted relationship where surface clues "hijack" or "dominate" it.

For ChatGPT-5 Mini, logistic regression analysis revealed that the context, the type of question, and their interaction terms all had high statistical significance ($p < 0.01$) (see Table 6). The odds ratio (OR) of the interaction term was as high as 15.521, indicating that the context effect was strongly modulated by the question type. Further analysis of the proportion of NP A attachment (see Table 7) revealed an unexpected pattern: when the question directly pointed to NP A (NP A Probe), the proportion of NP A attachment in the multi-reference context (31.250%) was significantly higher than that in the single-reference context (9.375%), suggesting a certain influence of the context. However, when the question pointed to NP B (such as "department") (NP B Probe), the model chose the proportion of NP A attachment to be exceptionally high (single-reference 59.375%, multi-reference 84.375%). This means that regardless of the cues provided by the context (multi-reference should ideally support the modification of NP A for distinction), once the probe question focuses on NP B, the model's response baseline is forcibly set to support an interpretation that is unreasonable in the given context (i.e., the relative clause modifies the distant NP B). This pattern indicates that ChatGPT's decision is not based on an assessment of the overall coherence of the discourse, but presents a rigid interaction "hijacked" by the wording of the question: although the context information is perceived, its power is distorted and used to reinforce the pre-determined, sometimes unreasonable, response tendency by the question type.

Table 6. Binary Logistic Regression Analysis for ChatGPT-5Mini on Attributive Modification Ambiguity Task

Predictor	B(Coefficient)	SE	Wald	*df*	P-value	OR
Context	1.380	0.461	23.175	1	0.003***	3.976
Probe Type	2.552	0.464	30.274	1	0.000***	12.828
Context*Probe Type	2.742	0.663	17.132	1	0.000***	15.521
Constant	-2.198	0.457	8.948	1	0.000***	0.111

Table 7. Proportion of NPA Attachment Judgments by ChatGPT-5Mini

Probe Type	Context	n	Proportion of NPA Attachment
NPA Probe	Single Referent	32	9.375%
	Multiple Referents	32	31.250%
NP B Probe	Single Referent	32	59.375%
	Multiple Referents	32	84.375%

DeepSeek-R1 exhibits a different distortion pattern. Its data analysis reveals that the contextual effect is significant (OR = 7.122), the effect of question type is also significant but in the opposite direction (OR = 0.372), and the interaction term between the two is not significant (p = 0.183) (see Table 8). This indicates that DeepSeek's decision-making seems to be more strictly dominated by the discourse context, but there is a problem type "deviation" that is in the opposite direction to the contextual effect. In terms of proportion (see Table 9), when the question points to NP A, the multi-reference context causes the proportion of NP A attachment to rise sharply from 9.375% to 56.250%, demonstrating strong contextual sensitivity. However, when the question points to NP B, the attachment proportion of NP A in the multi-reference context (25.000%) is still higher than that in the single-reference context (9.375%), which contradicts the "indication" of the question type (NP B). This suggests that DeepSeek seems to regard the contextual cues and question cues as two independent "commands" that need to be followed simultaneously, sometimes even in conflict, and performs a mechanical weighting or superposition, rather than an organic integration to determine "which modification relationship is more reasonable in the given context for the specific noun pointed to by the question". For example, in the context of "two policies and a department" (see Example 7), the model fails to, like humans, based on the discourse demand of "need to determine which of the two policies", stably bind the relative clause "that was outdated" with NP A (policy) and respond flexibly to different questions.

Example 7.

Multiple referents, NP A probe: There were two policies and a department. We compared the policy of the department that was outdated. Was the policy outdated, yes or no? [yes]

Multiple referents, NP B probe: There were two policies and a department. We compared the policy of the department that was outdated. Was the department outdated, yes or no? [yes]

Table 8. Binary Logistic Regression Analysis for DeepSeek-R1 on Attributive Modification Ambiguity Task

Predictor	B(Coefficient)	SE	Wald	*df*	P-value	OR
Context	1.963	0.510	14.799	1	0.000***	7.122
Probe Type	-0.999	0.459	4.639	1	0.031**	0.372
Context*Probe Type	-1.350	1.015	1.770	1	0.183	0.259
Constant	-1.872	0.455	16.902	1	0.000***	0.154

Table 9. Proportion of NPA Attachment Judgments by DeepSeek-R1

Probe Type	Context	n	Proportion of NPA Attachment
NPA Probe	Single Referent	32	9.375%
	Multiple Referents	32	56.250%
NP B Probe	Single Referent	32	9.375%
	Multiple Referents	32	25.000%

4.3.2 General Discussion: Limitations in Contextual Integration

The failure of resolving the attributive modification and PP attachment ambiguity is not an accident. They both point to a deeper core defect of the current AI systems based on Transformer architecture, represented by large language models: the lack of the ability to flexibly, rationally, and adaptively integrate multi-source information in a dynamic context (Cai et al., 2024; Yuan, 2024). Human sentence processing is widely understood as a constraint-based process in which multiple sources of information—syntactic structure, lexical semantics, discourse context, and world knowledge—are evaluated simultaneously to arrive at the most coherent interpretation. When resolving such structural ambiguities, the listener or reader can almost instantaneously and unconsciously weigh multiple constraints, including syntactic preferences, lexical semantic compatibility, the topic focus and referential clarity requirements of the discourse, and related world knowledge, ultimately integrating to form an optimal, internally coherent mental scenario model (Altmann & Steedman, 1988).

However, the performance of ChatGPT-5Mini and DeepSeek-R1 in this research reveals that, regardless of their architectural differences, neither of them was able to achieve such a level of integration. In both tasks, the performance of the models can be understood as their reliance on statistical shortcuts or heuristic strategies learned from the massive training data, specifically for combinations of surface cues (such as "with" + a specific verb/noun; specific interrogative words "who"/"that" + a specific noun position). For PP attachment, the models overly rely on the verb in the question (such as "use"); for attributive modification, the models are overly constrained by the noun mentioned in the question itself. Although context can be partially perceived and influence the output probability distribution by some models (such as DeepSeek), this influence is additive and mechanical, rather than integrative and inferential. The models do not seem to construct a unified discourse representation that can be dynamically updated as the discourse progresses and perform true "disambiguation"—that is, to evaluate and select the syntactic and semantic structure that makes the entire discourse context most coherent and reasonable (Linzen & Baroni, 2021).

These findings are consistent with previous research indicating that large language models can capture statistical regularities in language but may struggle with tasks requiring coordinated integration of multiple information sources (Cai et al., 2024; Linzen & Baroni, 2021). More broadly, it also echoes the concerns from the Wittgensteinian philosophical perspective: the meaning of language is deeply embedded in specific "life forms" and "language games". Current large language models may excel in imitating many "moves" in language games, but when it comes to understanding and responding to the key aspects of a discourse—for instance, understanding what exactly is being modified by "outdated" in "We compared the policy of the department that was outdated"—depending on the existence of several policies we already know—they still exhibit fundamental deficiencies. Therefore, to achieve human-level deep language understanding, future research may not merely rely on further expansion of model size and data, but instead needs to explicitly encourage and strengthen the dynamic and rational integration ability of cross-module and multi-level information in the architecture design or training goals.

5. Conclusion

This study has investigated the ability of large language models to process syntactic ambiguity from the perspective of both linguistic theory and empirical evaluation. Using a two-stage experimental design, it compared the performance of ChatGPT-5 Mini and DeepSeek-R1 in ambiguity detection and context-dependent disambiguation tasks. The findings show that both models perform above chance level in detecting syntactic ambiguity (DeepSeek-R1: 85% & ChatGPT-5 Mini: 77%), indicating that they can capture certain surface-level regularities associated with ambiguous structures. However, their performance remains below reported human benchmarks, and error analysis suggests that detection is less reliable when ambiguity cues are less prototypical or less frequent. This pattern points to a reliance on salient surface features rather than fully generalized structural representations. In the disambiguation tasks that require deep contextual integration, whether it is prepositional phrase attachment ambiguity or attributive modification ambiguity, both models failed to effectively utilize discourse information for reasonable resolution as humans do. Specifically, ChatGPT-5 Mini shows a strong dependence on surface-level cues, particularly the wording of probing questions, with minimal sensitivity to contextual variation while DeepSeek-R1 demonstrates partial responsiveness to contextual factors, but its behavior suggests a relatively shallow or additive combination of cues rather than fully integrated reasoning.

The above findings provide important empirical evidence for the current theoretical debate regarding whether large language models possess "true understanding capabilities". The results support the criticism made by Chomsky et al. (2023) that LLMs are "cumbersome statistical engines", indicating that the success of these systems is more due to the efficient capture and imitation of statistical patterns in massive data rather than a true simulation of the human understanding mechanism based on recursive syntactic rules and deep semantic integration. The systematic failures of the models in key tasks that require the coordinated operation of syntactic-semantic-pragmatic interfaces precisely confirm Quine's (2013) profound insight into the uncertainty of meaning and understanding—the understanding of language can never be fully reduced to statistical operations on surface symbols. At the same time, this study also reveals that the "emergent understanding" promised by the connectionist path (Couch, 2023) has not been fully realized at the current technological stage: even as the model size and data volume continue to expand, the cognitive ability to dynamically integrate multiple sources of information and construct a unified situational model does not seem to have emerged spontaneously (Linzen & Baroni, 2021).

Theoretically, this study provides an annotation from the era of artificial intelligence for the pragmatic view of "meaning is usage" emphasized by Wittgenstein (1953). The meaning of language is deeply rooted in specific "life forms" and "language games", and needs to be dynamically constructed based on shared contextual background and pragmatic principles. Although current large language models can excellently imitate many "moves" in language games, they still show fundamental deficiencies in understanding how discourse is embedded in specific contexts and how it serves specific communication purposes (Yuan, 2024). Practically, the findings of this study have important implications for the development of natural language processing applications. In scenarios such as machine translation, dialogue systems, and information extraction that require precise handling of complex syntactic structures, developers need to be aware of the limitations of current LLMs and avoid over-reliance on their surface fluency while ignoring potential risks of misunderstanding. At the same time, this study also calls for the establishment of a more refined, linguistically theory-driven evaluation paradigm to complement the traditional model evaluation system based on holistic indicators (Liu et al., 2023).

Several limitations of the present study should be acknowledged. Firstly, although the experimental corpus was systematically constructed, its scale is limited and it only focuses on a specific type of English syntactic ambiguity. The conclusion's generalization to other languages or broader ambiguity phenomena requires caution. Secondly, the model versions selected for the study (ChatGPT-5 Mini and DeepSeek-R1) represent the technical level at a specific time point. As the models evolve, their related capabilities may change dynamically. Future research can expand the scale of the corpus and the languages involved, incorporate more cutting-edge models for diachronic comparisons, and combine neuroscientific methods (such as brain imaging

techniques) to more directly compare the representations of the human brain and artificial neural networks in language processing (Caucheteux & King, 2022). Moreover, how to explicitly introduce constraints on structural dependence and cross-module integration in the model architecture or training objectives to promote the emergence of deeper understanding abilities is an important direction worth exploring (Levy, 2008).

In conclusion, through a systematic comparison of the sentence ambiguity handling capabilities of ChatGPT-5Mini and DeepSeek-R1, this study reveals the "integrative" understanding dilemma that exists behind the "fluency" of current large language models. This discovery not only provides an empirical response to the classic issues in language philosophy from the perspective of the artificial intelligence era, but also points out the next obstacle that needs to be overcome on the path to achieving deep language understanding at the human level.

Funding: This research received no external funding.
Conflicts of Interest: The author declares no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*(3), 191-238.https://doi.org/10.1016/0010-0277(88)90020-0
[2] Cai, Z., Duan, X., Haslett, D., Wang, S., & Pickering, M. (2024, August). Do large language models resemble humans in language use?. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 37-56). https://doi.org/10.18653/v1/2024.cmcl-1.4
[3] Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134. https://doi.org/10.1038/s42003-022-03036-1
[4] Carnie, A. (2021). *Syntax: A generative introduction*. John Wiley & Sons.
[5] Chomsky, N., Roberts, I., & Watumull, J. (2023). Noam chomsky: The false promise of chatgpt. *The New York Times*, *8*.
[6] Chomsky, N. (2002). *Syntactic structures*. Walter de Gruyter.
[7] Chomsky, N. (2014). *Aspects of the Theory of Syntax* (No. 11). MIT press.
[8] Chomsky, N. (2002). *On nature and language*. Cambridge University Press.
[9] Couch, J. R. (2023). Artificial intelligence: Past, present and future. *Journal of the South Carolina Academy of Science*, *21*(1), 2.
[10] Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, *2*(3), 5.
[11] Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177.https://doi.org/10.1016/j.cognition.2007.05.006
[12] Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195-212. https://doi.org/10.1146/annurev-linguistics-032020-051035
[13] Wittgenstein, L. (1953). Philosophical in-vestigations. *Trans: GEM Anscombe, Oxford: Blackwell*.
[14] Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., ... & Choi, Y. (2023). We're afraid language models aren't modeling ambiguity. *arXiv preprint arXiv:2304.14399*. https://doi.org/10.18653/v1/2023.emnlp-main.51
[15] Marcus, G. (2020). The next decade in AI: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.
[16] Piantadosi, S. T. (2023). Modern language models refute Chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, *15*, 353-414.
[17] Quine, W. V. O. (2013). *Word and object*. MIT press.
[18] Wittgenstein, L. (2009). *Philosophical investigations*. John Wiley & Sons.
[19] 袁毓林.(2024).ChatGPT语境下语言学的挑战和出路.现代外语,47(04),572-577.https://doi.org/10.20071/j.cnki.xdwy.20240523.009.
[20] 袁毓林.(2024).语义理解与常识推理的机器表现和人类基线之比较——怎样评估ChatGPT等大型语言模型的语言运用能力?.汉语学报,(04),2-16.
[21] 袁毓林.(2025).从三种复杂句看ChatGPT是不是随机鹦鹉?——语言大模型能不能理解语言意义的测试与讨论.语言教学与研究,(01),35-49.