Vol. 3 No. 3 (2025)

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895
https://llrjournal.com/index.php/11

Evaluating Ambiguity Resolution in DeepSeek: A Study on Lexical, Syntactic, and Semantic Disambiguation

Noor Ul Haya Khan
GIFT University, Gujranwala
Email: 211670175@gift.edu.pk

Arslan Ali
GIFT University, Gujranwala
Email: arslan.ali@gift.edu.pk

Faiza
University of the Punjab- Gujranwala Campus
Email: faiza.javed019@gmail.com

Javed Iqbal
II-TECH, Gujranwala
Email: javaidhaxat500@gmail.com

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

Abstract

Linguistic ambiguity has always challenged human communication, yet people resolve it effortlessly through context and background knowledge. For natural language processing systems (NLP), particularly in this era of large language models (LLMs), this fundamental aspect of language remains persistently problematic. As artificial intelligence and large language models (LLMs) continue to evolve, assessing their ability to interpret and resolve linguistic ambiguity has become increasingly important. This study examines how DeepSeek, an open-source LLM, deals with four types of linguistic ambiguity: homonymy, polysemy, syntactic, and semantic. A qualitative and descriptive approach is used in the study. It consists of a benchmark dataset of 33 test sentences. Prompts were entered into DeepSeek and the responses were manually examined and categorized into four groups: true positives, true negatives, false positives, and false negatives. The results show that DeepSeek performs well with syntactic and semantic ambiguity. It often provides detailed and grammatically accurate explanations. Nonetheless, the model found it hard to handle lexical ambiguity. It often over-detects ambiguity in homonymous sentences. Redundant alternative meanings of the polysemous phrases are proposed by the model even in obvious contexts. These results indicate that DeepSeek only comprehends the structure of the sentences used and fails to interpret the meaning through the context.

Keywords*:* Linguistic Ambiguity, Large Language Models (Llms), Deepseek, Natural Language Processing (Nlp), Lexical Ambiguity, Syntactic Ambiguity, Semantic Ambiguity

Introduction

Linguistic ambiguity has consistently represented an inherent feature of human language systems (Fortuny & Payrató, 2024). Human frequently reuse existing words in communication rather than inventing new ones for every concept. Consider the dual meaning of "*bank*," applying equally to financial institutions and riverbanks. Lexical ambiguity constitutes only part of the challenge. Syntactic structures generate their own complexities, as shown in the ambiguous sentence "I saw the man with the telescope," where the ownership of the telescope remains ambiguous between speaker and subject (Ghosh, 2025). Fortuny and Payrató (2024) explore how such issues arise not only in morphology, semantics and syntax. Humans usually resolve these kinds of ambiguities using context awareness. However, such tasks remain difficult for language models and computational systems. This reveals a significant gap between human understanding and machine processing.

Researchers have consistently highlighted how linguistic ambiguity can interfere with the effectiveness of natural language processing systems. Whether the goal is to translate text, respond to user queries, or manage dialogue, an NLP model must determine what a user actually means when a phrase can be interpreted in more than one way. According to Jusoh (2018), ambiguity is not just a minor inconvenience but a fundamental barrier to creating dependable language technologies. Yadav, Patel and Shah (2021) also point out that if a system can't handle vague or unclear language, it may slow down, misinterpret the input, or stop working properly. A big part of this

issue comes from the fact that machines don't naturally understand context the way humans do. Words with several meanings, or expressions tied to culture or situation, are usually fine for people but difficult for machines to process. Even well-known models like BERT still struggle with meaning that isn't explicitly spelled out in their training data (Abike & Honnry, 2025).

Jia, Morris, Ye, Sarro and Mechtaev (2025) point out that large language models sometimes ask follow-up questions that don't actually help clarify what the user meant. Such inefficiencies create verbose outputs and increase user cognitive load. Human ambiguity resolution relies on contextual and world knowledge (Bender & Koller, 2020), but computational systems depend solely on textual patterns. This limitation explains why even advanced models often fail to match human interpretive expectations (Lake & Murphy, 2023).

Contemporary language models including GPT-4, PaLM, Claude and Deepseek have demonstrated impressive fluency and coherence. This linguistic proficiency has sparked both excitement about their potential and careful analysis of their true understanding capabilities. However, despite their impressive fluency, many models still struggle to reliably identify which ambiguous phrases might cause real misunderstandings.

To ensure that language technologies produce reliable output for users, it is necessary for models to interpret input with a high degree of linguistic accuracy. Language technologies can only serve users well when they achieve precise text interpretation. As Qamar, Yasmeen, Pathak, Sohail, Madsen and Rangarajan (2024) demonstrates that moving beyond superficial processing to thorough linguistic analysis is essential for accurate results.

The rapid development of language models has intensified examination of their ambiguity resolution capabilities. When OpenAI launched ChatGPT in November 2022, it fundamentally changed how the public interacts with AI-generated content, creating massive mainstream interest. Later, the introduction of DeepSeek by the Chinese company Hangzhou DeepSeek AI Co. reignited discussions around the next generation of AI tools and their interpretive abilities (Mota, 2025).

DeepSeek's rapid rise has been marked by frequent model releases, including DeepSeek-Coder (November 2023), DeepSeek-MoE, and DeepSeek-V2 (May 2024). This rapid succession of models triggered intense competition among China's AI companies(Suryawanshi, 2025). Independent testing shows these models excel at multiple tasks including visual comprehension, question answering, and natural language generation (Ma, Zhao, Wang, Wang, Yuan, Chen, Bai, & Ren, 2025)

DeepSeek has emerged as a significant area of study in AI research, as findings show that it performs competitively against well-known models (Gao, Jin, Ke, & Moryoussef, 2025).Although DeepSeek is relatively new in public discourse yet academic investigations into its open-source versions have increased since 2023. This highlights the rising attention toward its architecture, functionalities, and real-world uses (Puspitasari, Zhang, Dam, Zhang, Kim, Hong, Bae, Qin, Wei, & Wang, 2025).Despite these advancements, independent linguistic studies have yet to thoroughly assess how effectively it interprets and handles ambiguity, which is a crucial aspect of true language understanding. This thesis addresses that gap by conducting a systematic analysis of DeepSeek's behavior in the face of different types of linguistic ambiguity: homonymy, polysemy, syntactic ambiguity, and semantic ambiguity.

DeepSeek-V3 is a sparse mixture-of-experts (MoE) language model that builds on recent advancements in efficiency and reasoning. the model uses expert routing and dynamic mechanisms to improve contextual understanding and language generation (Sands, Wang, Xu, Zhou, Wei, & Chandra, 2025).

The present research adopts a qualitative and descriptive approach. It opted for pre-validated linguistic stimuli sourced from a prior benchmark study by (Ortega-Martín, García-Sierra, Ardoiz, Álvarez, Armenteros, & Alonso, 2023). Rather than designing new test materials, the study adapts the original prompts to assess DeepSeek's responses. The methodology emphasizes both accuracy in ambiguity detection and the interpretive reasoning provided by the model.

This study aims to examine how effectively DeepSeek can recognize and address different types of ambiguity, as well as to identify common patterns of errors that occur during its language processing. This includes examining whether the model can appropriately distinguish between multiple possible meanings and whether it over-identifies or misses ambiguity in different contexts.

Research Questions

What are the most frequent error patterns in DeepSeek's handling of ambiguous language?

How effectively does the DeepSeek model interpret and resolve different types of linguistic ambiguity, including lexical (homonymy and polysemy), syntactic, and semantic ambiguity?

This research holds significance for both computational linguistics and applied NLP. By examining how a state-of-the-art model interprets linguistic ambiguity, the findings contribute to a deeper understanding of LLMs' semantic reasoning abilities and their limitations. It also highlights the areas in which model responses diverge from human interpretive expectations, thereby informing future model training and evaluation strategies.

The scope of this study is limited to four types of ambiguity, using a total of 33 test sentences categorized under homonymy, polysemy, syntactic ambiguity, and semantic ambiguity. The study does not aim to evaluate the full range of DeepSeek's linguistic capabilities but rather focuses on ambiguity as a specific, challenging phenomenon in language understanding.

Literature Review

Linguistics

Linguistics is the systematic study of language, encompassing the analysis of its structure, meaning, and use in various contexts. It also considers the cultural, social, historical, and political factors that influence language. Linguists generally analyze human language by exploring the connection between sound and meaning. According to Rao (2021),linguistics can be defined as "the science of language" or "the systematic study of language." As a theoretical discipline, linguistics aims to develop models that explain different aspects of language and contribute to a general theory of how language functions. Advancements in language teaching and learning often reflect the progression of linguistic theory. As the science of language, linguistics includes the study of phonology (sound systems), morphology (word formation), syntax (sentence structure), semantics (how meaning is conveyed), and the lexicon (mental vocabulary). Linguistics itself is further divided into several subdisciplines.

Phonetics, the scientific study of speech sounds, is essential for understanding how people produce, hear, and identify the sounds of language (Ladefoged & Johnson, 2015).**Phonology** is a linguistic branch that studies how speech sounds relate within and across languages (Gafni, 2025).Morphology examines the structure, classification, and function of the smallest meaning-bearing units in language, which are used to build words, phrases, and sentences. These units, known as **morphemes**, are combinations of sounds that carry either clear or implied meanings. Syntax explores the principles that determine the orderly arrangement of words into phrases, the combination of phrases into clauses, and the organization of clauses into well-formed sentences semantics is a core branch of linguistics that focuses on the study of meaning within language. (Dash, 2011).According to YULIANA (2023), it explores how meaning is structured and communicated through words, phrases, and sentences, and how it varies with context. Pragmatics is the study of how meaning and language use are shaped by the speaker, the listener, and contextual factors surrounding an utterance. It focuses on the influence of context in communication during speech events (Dash, 2011).

Beyond these core subfields, linguistics also intersects with technology through computational linguistics**.**

Computational Linguistics

Computational linguistics represents a modern and interdisciplinary branch of linguistics that merges concepts from both the arts and sciences. It plays a key role in the development of artificial intelligence within the field of language study. One of its most prominent subfields is Natural Language Processing (NLP), which applies computational techniques to analyze and simulate human language (Shokhrukh & Abror, 2022).

NLP

Building on the advancements in computational linguistics, Natural Language Processing (NLP) has emerged as a vital field that combines linguistic knowledge with artificial intelligence. It enables machines to understand, analyze, and generate language that resembles natural communication. NLP supports a wide range of modern technologies such as digital assistants, automated translation, and sentiment analysis and contributes to more seamless and intuitive human-computer interaction (Mulyadi; Saefudin, 2023).

A major milestone in NLP has been the emergence of large language models (LLMs), such as ChatGPT, GPT-4, Claude 3.5, Qwen, and the more recent DeepSeek. Among these, ChatGPT gained widespread recognition for its human-like conversational abilities. Whereas, DeepSeek which is considered the latest among these models, incorporates highly efficient and accurate algorithms. These tools excel in tasks like code generation, image description, and question answering However, they continue to face limitations in areas such as contextual understanding, bias, and reliability (Phogat, Arora, Mehra, Sharma, & Chawla, 2025).

While NLP focuses on enabling machines to process and understand human language but a key challenge is deciphering the true meaning behind words and sentences. This is particularly complex because language is often ambiguous and meaning can change based on context, tone, and sentence structure.

Ambiguity

Ambiguity occurs when a word, phrase, or expression can be interpreted in multiple ways, creating confusion or lack of clarity due to insufficient contextual clues (Elena, 2019). According to Larina, Ozyumenko and Ponton (2019), ambiguity often arises when a word or sentence allows for more than one interpretations, compelling the reader to choose between them. These layers of ambiguity are not merely theoretical curiosities; they play a critical role in shaping how meaning is communicated, influencing everything from casual conversations to academic writing. For instance, the sentence *"I saw the bank this morning"* contains a clear ambiguity. Without further context, the term 'bank' can signify a financial organization, the edge of a river, or even a blood bank. As noted in *the Longman Dictionary (2008),* each of these meanings exists independently, and understanding which one is intended relies heavily on context and timing

Types of Ambiguity

Linguistically, the phenomenon of ambiguity can be divided into different categories such as lexical, syntactic, semantic, pragmatic, and discourse ambiguity.

Ambiguity and Natural Language Processing (NLP)

As previously discussed, ambiguity is an inherent feature of language, present at various linguistic levels including lexical, syntactic, and semantic. It often leading to multiple interpretations of a word or sentence. Human communication easily handles ambiguity using context and shared knowledge, but this remains a major challenge for Natural Language Processing (NLP). As Ortega-Martín et al. (2023) note, human language often contains vague or context-dependent expressions that makes interpretation hard. While humans rely on intuition, machines struggle with these nuances which leads to errors in understanding.

Despite progress in NLP, ambiguity remains a key obstacle in NLP. Jusoh (2018) highlight that unclear meanings in language make it difficult for AI systems to process text accurately and consistently.

These systems often fail with idioms and context-heavy phrases. A frequently cited example is the sentence *"I saw him walking by the bank"* (Qamar et al., 2024) . This allows for multiple interpretations depending on whether "bank" refers to a financial institution or a riverbank, and whether "walking by" implies observation or accompaniment.

Tanjim, In, Chen, Bursztyn, Rossi, Kim, Ren, Muppala, Jiang and Kim (2025) note that ambiguity is still a major NLP challenge because human language is complex and flexible. With the growing use of Large Language Models (LLMs), solving ambiguity has become even more critical.

While all ambiguities challenge NLP systems, recognizing their different types is key for proper language understanding. Ambiguity can come from word choices, sentence structure, multiple meanings, or unclear references, with each type presenting distinct obstacles for computational systems.

Lexical ambiguity

Lexical ambiguity arises when a single word form can be interpreted in more than one way, depending on the context in which it is used (Ghosh, 2025).

Polysemy
Polysemy refers to cases where a single word has multiple related meanings. These meanings often evolve from metaphorical or contextual extensions. For example, the word "head" can indicate a body part, a leader, the top of a beverage, or the front position in a queue (Ghosh, 2025).

Homonymy
Homonymy occurs when two or more completely unrelated meanings share the same word form, either in spelling or pronunciation. For instance, the word *"bat"* may refer to a flying animal or a piece of sports equipment, while "left" can be understood as a direction or the past tense of leave (Ghosh, 2025).

Syntactic ambiguity
Syntactic ambiguity, also known as structural ambiguity, occurs when the grammatical arrangement of words in a sentence allows for more than one valid interpretation. The ambiguity arises not from the words themselves, but from the structure in which they are placed. For example, *"She hit the man with an umbrella"* may mean that she used an umbrella to hit the man, or that she hit a man who was holding an umbrella (Ghosh, 2025).

Semantic Ambiguity
Semantic ambiguity may arise in words, phrases, or sentences due to lexical, phonological, structural, grammatical, or semantic factors. This ambiguity may be introduced deliberately or happen unintentionally (Saleh, 2017).

Referential Ambiguity
Referential ambiguity is a subtype of semantic ambiguity that occurs when the intended referent of a pronoun or deictic expression (such as he, she, it, they, here, or there) is unclear. This type of ambiguity often arises when contextual information is insufficient to determine the referent, making interpretation challenging for both humans and NLP systems. For example, in the sentence “He told him that they would see it there,” the specific entities referred to by he, him, they, it, and there remain ambiguous without further context (Ghosh, 2025).
The recent NLP applications which rely on the basis of transformer frameworks, such as BERT and GPT showed quite substantial progress but are still facing linguistic ambiguity barriers. These problems are particularly common in the case of machine-translation. Even though, such systems introduce perceivable advancements in contextual meanings of words, they tend to fail when it comes to resolving polysemy and to interpret complex semantic roles. Such limitations are even more common in the cases where very different language systems are involved. Specifically, the weaknesses in the performance are evident in those languages that adopt divergent syntactic forms. According to Feng (2025) such issues hinder translation quality. These results indicate that current systems are not able to reach the human level of understandings about contexts.
Current studies have taken a keen interest in examining the way in which large language models (such as ChatGPT) confront various forms of lingual ambiguity. Such findings show that although ChatGPT can identify certain lexical ambiguities (e.g., polysemy), it often gives too large quantity of possible meanings, even when

they are not required. It happens because the model relies on word patterns and not on actual language understanding. The model also does not perform very well with semantic ambiguity, especially resolution of pronoun ambiguity. Although it is able to detect ambiguous pronouns, it often reverts to gender-based stereotypes rather than actually analyzing the situation. These limitations, according to researchers, can be explained by the fact that the system applies statistical word patterns instead of actual comprehension of meaning (Ortega-Martín et al., 2023).

Taking it one step further, Qamar et al. (2024) aimed at determining whether ChatGPT can handle different forms of language ambiguity. Wordplay, code-mixing, lexical, syntactic, semantic ambiguity were analyzed. When easy or textbook examples were used (such as traditional examples of polysemy), the model performed fine. However, on a more advanced or real-life language, it was recorded that the results were weaker. For example, ChatGPT was not able to deduce the meaning of a code-mixed word “doctaron”, which is a blend of Hindi and English. It failed to understand figurative meanings as well. When given the word “doormat,” it interpreted it literally and did not recognize it as a metaphor for a submissive person. it also failed to resolve riddles or puns (e.g., identifying "frog" in "If Roger comes"). These limitations occur because the model relies too much on the training data. It often fails to understand creative or unusual uses of language.

Mulyadi; Saefudin (2023) demonstrates ChatGPT's sensitivity to input phrasing and shows that the model often misinterprets ambiguous terms unless explicitly disambiguated by the user, highlighting limitations in handling linguistic variations.

Collectively, these studies underscore LLMs’ reliance on pattern recognition rather than grounded linguistic reasoning, exposing a critical need for architectures robust to ambiguity.

Research Gap

While existing studies have analyzed how ChatGPT deals with different types of linguistic ambiguity including lexical, syntactic, and semantic domains. However, there hasn’t been much work on newer language models like DeepSeek. Even though DeepSeek is becoming more well-known and uses its own training methods, no study so far has tested how it handles ambiguous language cases, particularly tasks involving syntactic disambiguation, lexical ambiguity resolution, metaphor interpretation, and co-reference resolution. This creates a critical gap in the literature, as understanding how newer models like DeepSeek process ambiguity is essential for benchmarking progress beyond ChatGPT and for evaluating whether recent architectural shifts address the limitations previously identified. Therefore, this study aims to investigate how DeepSeek performs in resolving ambiguity, contributing to a broader understanding of LLMs’ language competence and robustness.

Research Methodology

Research Design

This research follows a qualitative and descriptive aimed at evaluating the ability of the DeepSeek language model to interpret and resolve various types of linguistic ambiguity.

Qualitative research is defined as "an interpretive, naturalistic approach to the world, meaning that researchers study things in their natural settings, attempting to make sense of phenomena in terms of the meanings people bring to them" (Denzin, Lincoln,

MacLure, Otterstad, Torrance, Cannella, Koro-Ljungberg, & McTier, 2017). Following this framework, the study analyzes DeepSeek's outputs to identify reasoning patterns, linguistic behaviors, and limitations in handling ambiguity.

Descriptive research, as described by Babbie (2020), involves efforts "to document the current state of affairs, providing a snapshot of conditions, attitudes, or behaviors." The aim here is to observe how DeepSeek responds linguistic structures, and to organize those responses into different categories. No experimental manipulation is involved. Instead, the model's responses are classified using interpretive categories (e.g., true positive, false negative) to offer a systematic account of its ambiguity-handling capabilities.

The study builds upon an established benchmark by replicating the experimental framework used in Ortega-Martín et al. (2023). The researchers preferred to use linguistic stimuli that had been previously validated instead of using new items to protect the reliability of the study.

Data Collection

The research uses the ambiguity detection dataset from Ortega-Martín et al. (2023) with little changes. There are instances of ambiguous and non-ambiguous sentences in the dataset which are categorized into four domains of homonymy, polysemy, syntactic ambiguity and semantic ambiguity. Specifically:

A total of 33 sentences were used for evaluation.

These included:

13 sentences under homonymy

10 sentences under polysemy

6 sentences under syntactic ambiguity

4 sentences under semantic ambiguity

Each of these input sentences was fed to DeepSeek-V3, a large-scale generative artificial-intelligence model focused on natural-language understanding and generation. The model has a publicly available user interface through which all the inputs were manually entered into the model.

These sentences were analyzed in the original study and were known to represent certain types of ambiguity. Thus, the need to develop new test items was minimized.

Tools

The same prompt was inserted into the DeepSeek language model for each sentence: "Is the sentence "[sentence]" ambiguous?"

An additional question was included in the case of syntactic ambiguity, which was concerned with ambiguity at clausal level. This latter version was called Prompt 2 and consisted of the following format:

"In the sentence "[sentence]", is the clause '[clause]' ambiguous?"

Prompt 2 was designed check the knowledge of DeepSeek about the structure of complex sentences, more specifically on the grounds of its ability to detect ambiguities of clause- level attachments.

Data Analysis

After every sentence was fed into the model individually, responses were gathered to undergo a systematic analysis. All the inputs were of the following simple format:

"Is the sentence '[sentence]' ambiguous?"

This was the significant prompt model used to gauge the performance of this model on four major categories of ambiguity: homonymy, polysemy, syntactic, and semantic. A second prompt structure was used to measure clause-specific interpretation of syntactically ambiguous phrases. The second prompt was based on the following structural schema: “In the sentence '[sentence]', is the clause '[clause]' ambiguous?”. The extra approach allowed making a more precise estimate of the ability of the model to analyze sentence structure and identify the existence of clause-level ambiguity.

The manual assessment of each DeepSeek output was performed and responses were classified into one of the four outcomes as shown in the list below:

True Positive (TP): The model accurately detected an ambiguous sentence.

True Negative (TN): The model accurately detected a non-ambiguous sentence.

False Positive (FP): The model incorrectly labeled a non-ambiguous sentence as ambiguous.

False Negative (FN): The model failed to detect an actual ambiguous sentence.

The current study conducted a qualitative assessment of the model losses of DeepSeek.

This included:

An assessment of output to determine whether the model is offering lexical, syntactic or semantic descriptions in the output.

Assessing whether the kind of ambiguity model identified was possible in practice or regarded as being merely theoretical.

General inspection of any trends in the behavior of the model with various kinds of ambiguity.

The number of the true positives, false positives, true negatives as well as the false negatives was listed in the summary, and quantitative findings for each category of ambiguity were arranged in cumulative tables. Overall patterns and more general trends in ambiguity categories were discovered through these counts.

Results

This section gives an account of how DeepSeek processes various types of ambiguity. These are homonymy, polysemy, syntactic and semantic ambiguity. The explanations of the answers given by the model are provided with the examples from the dataset and every subsection contains a summary of the model's behavior.

The ability of DeepSeek to identify and overcome ambiguity is tested empirically in this study. Model responses were organized according to the type of ambiguity present in each sentence and were further classified into four outcome categories based on their accuracy:

True Positive (TP): The model correctly identified an ambiguous sentence.

True Negative (TN): The model correctly rejected a non-ambiguous sentence.

False Positive (FP): The model incorrectly labeled a non-ambiguous sentence as ambiguous.

False Negative (FN): The model failed to recognize an ambiguous sentence.

Each category is discussed in detail below:

Homonymy

DeepSeek was presented with 13 sentences containing homonymous word pairs. Of these, only 3 were genuinely ambiguous, yet the model labeled all 13 as ambiguous

While each sentence did include homonyms (words with different, unrelated meanings), many of them were contextually clear and would not typically be considered ambiguous by human readers. This led to a high number of false positives, suggesting DeepSeek over-identifies ambiguity in homonymous constructions.

Table 1
DeepSeek's Sentence-Level Performance on Homonymy Prompts

No.	Sentence	Ambiguity Detected	Verdict	DeepSeek Verdict Summary
1	How does a bear bear with the pain?	Yes	FP	Recognized noun vs. verb homonymy; labeled as ambiguous.
2	The man bought a chocolate bar at the bar.	Yes	FP	Identified place vs. object meanings; labeled as ambiguous.
3	Going to the right is the right choice.	Yes	FP	Identified directional vs. moral homonymy; labeled as ambiguous.
4	There is no choice left but going to the left.	Yes	FP	Labeled ambiguous; interpreted both residual and directional senses.
5	The pain in my back came back after work.	Yes	FP	Labeled technically ambiguous; noted unnatural reading.
6	The band singer was wearing a head band during concert.	Yes	FP	Overanalyzed and identified dual meanings of "band"; labeled ambiguous.
7	You should address the problem with your address.	Yes	FP	Detected verb vs. noun homonymy; labeled as ambiguous.
8	Rose sale rose during February.	Yes	FP	Identified noun vs. verb homonymy; labeled as ambiguous.
9	Did you watch the watch my mom gave me?	Yes	FP	Identified verb vs. noun senses of "watch"; labeled ambiguous.
10	Can you close the trash can?	Yes	FP	Labeled as ambiguous due to "close" (shut vs. bring near) Misidentified "close" as source; ignored "can" (noun vs modal).
11	I bought a bar	Yes	TP	Labeled as ambiguous; Identifies multiple unrelated meanings: pub, chocolate bar, metal bar, legal bar
12	You look right	Yes	TP	Labeled as ambiguous; Describes both directional ("look to the right") and evaluative ("you look appropriate") interpretations
13	I saw a band	Yes	TP	Recognizes homonyms: music group, material strip, rebel group, technical frequency band; Labeled as ambiguous

For instance, in the sentence *"The man bought a chocolate bar at the bar,"* DeepSeek labeled this as ambiguous, pointing out that "bar" could refer to a chocolate bar (a rectangular piece of food) or a bar as in a drinking establishment. While both meanings exist, in this sentence, the second instance of "bar" clearly refers to a location (the pub), and the first to the object (the chocolate bar). There is no confusion or overlap; the sentence is structurally and contextually unambiguous Similarly, in "You should address the problem with your address," the model identified the dual usage of ″address″ as both a verb and a noun, labeling the sentence as ambiguous. While the first instance is a verb meaning *to* deal with, and the second is a noun referring to a location, the sentence poses no real ambiguity

In contrast, a sentence like "I bought a bar" genuinely lacks sufficient context to

determine whether "bar" refers to a metal rod, a pub, or a chocolate bar. Here, DeepSeek's ambiguous classification is justified, as multiple meanings remain plausible without further clarification.

Table 2
Summary of DeepSeek's Performance on Homonymy

Ambiguity Type	Total Prompts	True Positives	True Negatives	False Positives	False Negatives
Homonymy	13	3	0	10	0

Overall, DeepSeek performs inconsistently in handling homonymy. While it is effective at identifying when a word has multiple unrelated meanings (i.e., homonyms), it often fails to judge whether these meanings actually lead to ambiguity in context. As a result, it tends to over-label clear sentences as ambiguous, leading to high false positive rates and raising concerns about its reliability in real-world disambiguation tasks.

Polysemy
In the polysemy test, ten polysemous lemmas were tested in clearly non-ambiguous contexts to evaluate whether DeepSeek identifies them as ambiguous. The sentences were constructed to activate only one dominant sense of the lemma in context, ensuring that while the word had multiple related meanings, the intended one was obvious.
DeepSeek performed **exceptionally well** on this task. It correctly labeled all 10 sentences as **non-ambiguous**, resulting in a perfect **True Negative** score. The model shows appropriate sensitivity to polysemy and does not over-detect ambiguity in these examples.

Table 3
DeepSeek's Sentence-Level Performance on Polysemy Prompts

No	Lemma	Sentence	Ambiguity detected	Verdict	DeepSeek Verdict Summary
1	Serve	She has served on the committee for the last 15 years.	No	TN	Not ambiguous; discusses tense (ongoing or ended service).
2	Big	The house has four bedrooms, so it's pretty big.	No	TN	Not ambiguous; points out subjectivity of "pretty big," but accepts common understanding.
3	Blocked	The road was completely blocked by an overturned truck.	No	TN	Clear and direct; no ambiguity detected.
4	Glass	A glass of orange juice.	No	TN	Not ambiguous; notes far-fetched overinterpretations but confirms clarity.

5	Cheat	She found out that he'd been cheating on her.	No	TN	Not ambiguous; explored emotional vs. potential financial cheating but leaned on primary romantic meaning.
6	Scan	A laser beam scans the disc's surface.	No	TN	Clear literal meaning; acknowledged missing technical details but rejected any ambiguity.
7	Mouse	Once you get the feel of it, using a mouse is easy.	No	TN	Mild ambiguity; noted possible confusion between device vs. animal but leaned toward computer mouse.
8	Run	He's been running a restaurant since he left school.	No	TN	Not ambiguous; clearly interpreted as managing a business. Rejected alternate humorous readings.
9	Date	He asked her out on a date.	No	TN	Not ambiguous; identified romantic meaning as dominant. Explored alternate readings only hypothetically.
10	Bank	By the time we reached the opposite bank, the boat was sinking.	No	TN	Not ambiguous; considered possible confusion due to lack of river context but confirmed clarity of standard meaning.

Importantly, DeepSeek not only labeled these sentences but also provided detailed interpretations, often offering multiple possible senses of the lemma and clarifying which ones were contextually likely or unlikely. For example, in the sentence "He asked her out on a date," the model explored alternative definitions of "date" (such as a calendar day or a romantic appointment), but ultimately confirmed the dominant reading as romantic. Similarly, in the sentence "Using a mouse is easy once you get the feel of it," it acknowledged both the animal and computer device interpretations, while rightly identifying the latter as contextually appropriate.

These results suggest that DeepSeek is **highly reliable when handling polysemous constructions**, as it consistently uses **contextual cues** to disambiguate related word meanings. Here, it demonstrated strong pragmatic awareness and an ability to suppress unnecessary interpretations.

Table 4

Summary of DeepSeek's Performance on Polysemy

Ambiguity Type	Total Prompts	True Positives	True Negatives	False Positives	False Negatives
Polysemy	10	0	10	0	0

Although DeepSeek correctly identifies certain sentences as unambiguous, it often goes on to suggest additional interpretations that are implausible or unlikely from a human perspective. These alternative readings, while theoretically possible, do not align with how the sentence would be understood in natural communication. This tendency to list multiple meanings, even after confirming the sentence is unambiguous, may introduce confusion and undermine the clarity of the model's judgment.

Syntactic

For syntactic ambiguity, DeepSeek exhibited notably accurate performance. A total of 5 distinct ambiguous constructions were tested, with some items repeated under Prompt 2 to verify consistency and depth of explanation. In every case, the model correctly identified the ambiguity and provided insightful, structurally grounded justifications. This results in 5 true positives, and no false positives or false negatives—indicating complete success in this category.

Table 5

DeepSeek's Sentence-Level Performance on Syntactic Ambiguity Prompts

No.	Sentence	Ambiguity Detected	Verdict	DeepSeek Verdict Summary
1	Mary saw John with a telescope	Yes	TP	Detailed and accurate explanation of PP-attachment ambiguity.
2	I am happy you are proud, and so is Mary	Yes	TP	Thorough analysis of clause coordination ambiguity.
3	I watched her duck	Yes	TP	Clear identification of lexical and syntactic dual reading; well-reasoned.
4	I saw the dog with one eye	Yes	TP	Accurately parsed as PP-attachment ambiguity.
5	They are cooking apples	Yes	TP	Correctly identifies Gerund ambiguity.
6	***Prompt 2*** *1)* Mary saw John with a telescope *2)* I am happy you are proud, and so is Mary	Yes	TP	Reaffirms prior ambiguities: explains attachment scope in (1) and elliptical reference in (2).

For instance, in the classic prepositional phrase (PP) attachment example **"Mary saw John with a telescope,"** DeepSeek correctly recognized that the ambiguity stems from whether "with a telescope" attaches to the verb phrase (Mary used the telescope)

or the noun phrase (John had the telescope). It provided both readings clearly and explained the syntactic structure that leads to each interpretation.

Similarly, the sentence ″I watched her duck″ was interpreted as syntactically ambiguous between a verb phrase (her quickly moving) and a noun phrase (her pet duck). DeepSeek successfully outlined the structural duality and acknowledged that disambiguation depends on context.

Further, when tested with **Prompt 2—**which reintroduced two syntactically ambiguous sentences—DeepSeek maintained consistency in its responses. It not only repeated the correct identification of ambiguity but expanded its explanation. the model analyzed the sentence "I am happy you are proud, and so is Mary," and successfully identified two sources of linguistic ambiguity: ellipsis **and coordination**. It correctly determined that the phrase "so is Mary" could refer either to "being happy" or to "being proud." The findings confirm that the model is capable of advanced syntactic analysis way beyond the surface understanding.

Table 6
Summary of DeepSeek's Performance on Syntactic Ambiguity

Ambiguity Type	Total Prompts	True Positives	True Negatives	False Positives	False Negatives
Syntactic	5	5	0	0	0

DeepSeek provides in-depth syntactic explanations. The system precisely realizes the ambiguity and provides reasonable grammatical explanations for the phenomena like clauses attachment, ellipsis and coordination. It is worth noting that the model addresses the concept of syntactic ambiguity on a quite detailed level as it examines the structure of sentences thoroughly and identifies areas where different word order lead to syntactic ambiguity.

Semantic

DeepSeek performed well in identifying semantic ambiguity. Its strongest responses appeared in cases involving pronoun reference and co-reference resolution. Out of four test items, the model correctly detected ambiguity in three. In each of these cases, it gave explanations that reflected a clear understanding of the underlying meaning. One response, however, was incomplete and resulted in a false negative.

Table 7
DeepSeek's Sentence-Level Performance on Semantic Ambiguity Prompts

No.	Sentence	Ambiguity Detected	Verdict	DeepSeek Verdict Summary
1	My mother and my sister were sad after she shouted at her	Yes	TP	Clear and precise explanation of co-reference ambiguity.
2	The physicist hired the secretary because she was overwhelmed with work	Yes	TP	Thorough interpretation of pronoun reference ambiguity.
3	The physicist hired the secretary because he was overwhelmed with work	Yes	TP	Clear explanation with attention to gender cues in pronoun reference

				ambiguity.
4	"The physicist hired the secretary because she was overwhelmed with work." who the pronoun "he" refers to?", who the pronoun "she" refers to?	partial	TP	DeepSeek ignored the query about "he" despite its inclusion in the prompt.

For instance, in the sentence "The physicist hired the secretary because she was overwhelmed with work," DeepSeek identified that "she" could refer either to the physicist or the secretary. Importantly, it noted that while societal stereotypes might suggest the physicist is male and the secretary is female, this interpretation is not grammatically guaranteed. The model emphasized grammatical structure over gender-based assumptions, showing that its interpretation was driven by syntactic cues rather than social bias.

However, in the related sentence "The physicist hired the secretary because she was overwhelmed with work. Who does 'he' refer to?", DeepSeek only addressed the ambiguity of "she" and ignored the part of the question concerning "he". Although "he" was not explicitly mentioned in the sentence, its inclusion in the follow-up query implies a contrastive co-reference resolution, if one individual is "she," the other could be assumed to be "he." DeepSeek's failure to respond to this element resulted in a missed opportunity to resolve a more complex referential structure, contributing to its lone false negative in this category.

Table 8
Summary of DeepSeek's Performance on Semantic Ambiguity

Ambiguity Type	Total Prompts	True Positives	True Negatives	False Positives	False Negatives
Semantic	4	3	0	0	1

Despite this, DeepSeek consistently demonstrated a strong grasp of semantic ambiguity, particularly in parsing sentences where pronoun reference is contextually dependent. The model's responses reflected sound reasoning. It was able to distinguish grammatical ambiguity from interpretative bias. This distinction was particularly important when analyzing sentences that allowed for more than one possible reading.

Summary of Results

The following table summarizes DeepSeek's performance across all four types of linguistic ambiguity tested in this study.

Table 9
Performance of Deepseek LLM Across Linguistic Ambiguity Types

Ambiguity Type	Total Prompts	True Positives	True Negatives	False Positives	False Negatives
Homonymy	13	3	0	10	0
Polysemy	10	0	10	0	0

Syntactic	6	6	0	0	0
Semantic	4	3	0	0	1[1]
TOTAL	**33**	**12**	**10**	**10**	**1**

DeepSeek achieved perfect accuracy in handling both polysemy and syntactic ambiguity. It reliably rejected false cases of ambiguity and correctly interpreted structurally ambiguous constructions. DeepSeek performed quite well on semantic ambiguity. Three of the four cases were accurately identified. However, it missed a case in which a pronoun reference was unclear. DeepSeek performed quite well on semantic ambiguity. Three of the four cases were accurately identified. However, it missed a case in which a pronoun reference was unclear. Its homonymy outcomes are much weaker. The model has a significant false positive rate because it misclassified all 10 sentences as ambiguous.

Conclusion

The purpose of this study was to assess whether the DeepSeek language model is able to understand and detect linguistic ambiguity. There are four types of ambiguity (Homonymy, polysemy, syntactic ambiguity, and semantic ambiguity) upon which the performance of the model was studied. Special emphasis was laid on the effectiveness of resolving ambiguity as well as on the most common error patterns that emerged during the analysis.

The findings show that DeepSeek overidentifies ambiguity in homonymous sentences. The model identifies homonymous constructions as ambiguous on numerous occasions even in situations where the context provides enough information to effectively clarify the intended meaning. Therefore, a large number of false positives is found in the homonymy category. In the case of polysemy, the model precisely identifies unambiguous sentences, but also derives unrelated alternative meanings in the result. These responses show that there is an advanced lexical understanding of the language model but on the same note they also show how the model has potential to mix two different word senses with true ambiguity.

DeepSeek proved to be effective in the field of syntactic and semantic ambiguity resolution. In regards to syntactic ambiguity, the system performed well in identifying ambiguous clauses and recording a correct analysis of underlying grammatical errors. It also managed semantic ambiguity fairly well with most successful co-reference resolution. However, the model had sometimes overlooked delicate references in complex pronoun structures. the system has shown powerful linguistic awareness when it comes to both syntactic and semantic ambiguity. However, it handled potential meaning variations a as actual ambiguity and the system is also prone to overgeneralization on lexical contexts. Such results mark a mixed performance and indicate that model's performance relies on the type of ambiguity it encounters. Based on these observations, it can be assumed that the model can be improved in its contextual judgment and explicit distinction of lexical variety and real ambiguity.

References

Abike, S., & Honnry, E. (2025). Latest Progress in NLP and Deepening Language Understanding.

Babbie, E. R. (2020). The practice of social research. Cengage Au.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of the 58th annual meeting of the association for computational linguistics,

Dash, N. S. (2011). Language and Linguistics: A Scientific Approach to Language. Heritage Publishers.
https://www.researchgate.net/publication/383649963_Language_and_Linguistics_A_Scientific_Approach_to_Language#fullTextFileContent

Denzin, N. K., Lincoln, Y. S., MacLure, M., Otterstad, A. M., Torrance, H., Cannella, G. S., Koro-Ljungberg, M., & McTier, T. (2017). Critical qualitative methodologies: Reconceptualizations and emergent construction. International Review of Qualitative Research, 10(4), 482-498.

Elena, B. (2019). Ambiguity matters in linguistics and translation. Слово. ру: Балтийский акцент, 10(3), 81-93.

Feng, C. (2025). Analysis of Semantic Deviation in AI Translation and Linguistic Optimization Paths. International Scientific Technical and Economic Research, 3(2). https://doi.org/10.71451/ISTAER2522

Fortuny, J., & Payrató, L. (2024). Ambiguity in linguistics 1. Studia Linguistica, 78(1), 1-7.

Gafni, C. (2025). Phonetics and Phonology.

Gao, T., Jin, J., Ke, Z. T., & Moryoussef, G. (2025). A comparison of deepseek and other LLMs. arXiv preprint arXiv:2502.03688.

Ghosh, D. P. (2025). Ambiguity Across Human and Machine Languages: Emergence, Context, and Tooling. In L. T. C. M. M. K. Engineering Design & Research Center, India (Ed.).

Jia, H., Morris, R., Ye, H., Sarro, F., & Mechtaev, S. (2025). Automated Repair of Ambiguous Natural Language Requirements. arXiv preprint arXiv:2505.07270.

Jusoh, S. (2018). A study on NLP applications and ambiguity problems. Journal of Theoretical & Applied Information Technology, 96(6).

Ladefoged, P., & Johnson, K. (2015). A Course in Phonetics (Seventh). Cengage Learning.

Lake, B. M., & Murphy, G. L. (2023). Word meaning in minds and machines. Psychological review, 130(2), 401.

Larina, T., Ozyumenko, V., & Ponton, D. M. (2019). Persuasion strategies in media discourse about Russia: Linguistic ambiguity and uncertainty. Lodz papers in Pragmatics, 15(1), 3-22.

Ma, B., Zhao, Y., Wang, J., Wang, G., Yuan, K., Chen, T., Bai, L., & Ren, H. (2025). Can DeepSeek Reason Like a Surgeon? An Empirical Evaluation for Vision-Language Understanding in Robotic-Assisted Surgery. arXiv preprint arXiv:2503.23130.

Mota, R. (2025, April 2025). Is DeepSeek a Metacognition AI? Retrieved July 29, 2025 from https://doi.org/10.32388/PJ3POM

Mulyadi; Saefudin, D. P. P., Shadam Hussaeni Handi. (2023). Linguistic

variations in ChatGPT. Philosophica: Jurnal Bahasa, Sastra, dan Budaya, 6(2). https://doi.org/10.35473/pho.v6i2.2715

Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Álvarez, J., Armenteros, J. C., & Alonso, A. (2023). Linguistic ambiguity analysis in ChatGPT. arXiv preprint arXiv:2302.06426.

Phogat, R., Arora, D., Mehra, P. S., Sharma, J., & Chawla, D. (2025). A Comparative Study of Large Language Models: ChatGPT, DeepSeek, Claude and Qwen. 2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT),

Puspitasari, F. D., Zhang, C., Dam, S. K., Zhang, M., Kim, T.-H., Hong, C. S., Bae, S.-H., Qin, C., Wei, J., & Wang, G. (2025). Deepseek models: A comprehensive survey of methods and applications. Authorea Preprints.

Qamar, M. T., Yasmeen, J., Pathak, S. K., Sohail, S. S., Madsen, D. Ø., & Rangarajan, M. (2024). Big claims, low outcomes: fact checking ChatGPT's efficacy in handling linguistic creativity and ambiguity. Cogent arts & humanities, 11(1), 2353984.

Rao, V. (2021). Cognitive linguistics: An approach to the study of language and thought. Cognitive Linguistics: An Approach to the Study of Language and Thought.

Saleh, Y. M. (2017). Semantic ambiguity In English language. Iraq: University of Samarra.

Sands, B., Wang, Y., Xu, C., Zhou, Y., Wei, L., & Chandra, R. (2025). An evaluation of LLMs for generating movie reviews: GPT-4o, Gemini-2.0 and DeepSeek-V3. arXiv preprint arXiv:2506.00312.

Shokhrukh, J. B., & Abror, S. H. (2022). THE DEVELOPMENT TENDENCIES OF COMPUTATIONAL LINGUISTICS IN UZBEKISTAN: NLP, MACHINE TRANSLATION, CORPUS LINGUISTICS AND AUTOMATIC TEXT EDITING. The American Journal of Social Science and Education Innovations, 4(10), 01-05.

Suryawanshi, S. M., Smita Vasantrao; Rathod, Pooja; Pawar, Digvijaysinh Anil. (2025). DeepSeek AI in Libraries. International Journal of Applied Ethics, 11(1). https://www.researchgate.net/publication/389215005_Deepseek_AI_in_Libraries#full TextFileContent

Tanjim, M. M., In, Y., Chen, X., Bursztyn, V. S., Rossi, R. A., Kim, S., Ren, G.-J., Muppala, V., Jiang, S., & Kim, Y. (2025). Disambiguation in Conversational Question Answering in the Era of LLM: A Survey. arXiv preprint arXiv:2505.12543.

Yadav, A., Patel, A., & Shah, M. (2021). A comprehensive review on resolving ambiguities in natural language processing. AI Open, 2, 85-92.

YULIANA, E. P. (2023). SEMANTIC ANALYSIS IN THE LYRICS OF COLDPLAY'S SONGS UIN RADEN INTAN LAMPUNG].