A Comparative Evaluation of Semantic Ambiguity Detection in Two LLMs

Lili Tamás

Károli Gáspár University of the Reformed Church in Hungary
Reviczky u. 4. 1088 Budapest, Hungary
lili.tamas.contact@gmail.com

Abstract

The growing popularity and misconceptions about conversational AI systems are driving efforts to establish a universally accepted framework for evaluating large language models. Testing large language models on tasks designed to assess human cognitive skills has become widespread. This paper presents the results of a pilot experiment and a comparative evaluation of the ability of OpenAI's GPT-4.1 and GPT-4.1 mini to detect semantic ambiguity based on the works of Shultz and Pilon (1973) and Zipke et al. (2009). The experiment used a task sheet of 116 items utilising riddles, single sentences, and sentence pairs. It included systematically varied instructions on a four-level scale ranging from no mention of ambiguity to direct mention. Lexical and structural ambiguity were both employed, including surface-structure and deep-structure ambiguity. The results suggest that even advanced models, such as GPT-4.1 and GPT-4.1 mini, tend to consider only one possible meaning of ambiguous sentences. However, the recognition of ambiguity improved quickly when the possibility of ambiguity was explicitly referenced in the instruction. Additionally, the results imply that model size is not directly connected to performance, as GPT-4.1 scored better on lexical ambiguity detection tasks, while GPT-4.1 mini surpassed the larger model in structural ambiguity detection.

Keywords: LLM evaluation, semantic ambiguity, contextual limitations, language modelling

1. Introduction

The rapid advancement of large language models imposes new challenges for performance evaluation. Compared to benchmarking in evaluating discriminative language models, assessing the performance of generative language models poses new challenges. As of now, there is no universally accepted framework for evaluating such models (Wolters et al., 2024; Tam et al., 2024; Seo et al., 2024; Miaschi et al., 2024). Methods that simulate tests that were initially intended to assess human cognitive skills have become widespread. One topic of interest is LLMs' ability to detect ambiguity and whether these systems demonstrate metalinguistic awareness. This report presents the results of a small-scale experiment and the comparative evaluation of the ability of OpenAI's GPT-4.1 and GPT-4.1 mini to detect semantic ambiguity.

Assessing whether large language models exhibit metalinguistic awareness is of utmost importance, as the general belief is that such models have a deep linguistic understanding of languages, beyond that of native speakers. As Rohr-Brackin (2025) explains, metalinguistic awareness is "a part of general cognition" (2025, p. 28). Metalinguistic awareness is the active attention to the knowledge domain "that describes the explicit properties of language" (Bialystok, 2021, qtd. in Roehr-Brackin, 2025, p. 28). Illiteracy and metalinguistic awareness are connected, as "illiterate adults' metalinguistic awareness remains at low levels, such as the ability to identify rhymes … despite cognitive maturity" (Roehr-Brackin, 2025, p. 29). Recognition of ambiguity requires metalinguistic awareness, and as such, it was chosen as the domain of the present research.

While the challenges of semantic ambiguity detection in NLP have been the focus of research for decades, Jayaweera and Dorr (2025) state that annotator-disagreement stemming from linguistic ambiguity is still often considered noise rather than a reflection of "meaningful, coexisting interpretations" (p. 37). Jayaweera and Dorr (2025) highlight that the "absence of gold-standard annotations for different ambiguity types hinders progress in training and evaluating models that aim to align more closely with human interpretive processes" (p. 44). The authors emphasise the "need for the creation of new datasets specifically annotated for ambiguity presence and type" and see "exploring unsupervised or weakly supervised methods" as promising (Jayaweera & Dorr, 2025, p. 44). The experiment presented in this current paper is to serve as a pilot project for a larger scale experiment as a step towards achieving the goals defined by Jayaweera and Dorr (2025).

2. Methods

The experiment tested the abilities of GPT-4.1 and GPT4.1 mini. OpenAI defines GPT-4.1 as their "smartest non-reasoning model", while GPT-4.1 mini is the "smaller, faster version of GPT-4.1" (OpenAI, Inc., n.d.). OpenAI does not share the detailed technical specifications of its models, but both models have a 1,047,576-token context window (OpenAI, Inc., n.d.). The tasks

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 60–65

used in the experiment were based on Shultz and Pilon's (1973) and Zipke et al.'s (2009) work on testing children's ability to detect and comprehend linguistic ambiguity. Each task consisted of a context and an instruction. As shown in Table 1, three types of contexts were used.

Context ID	Original study	Task type	Ambiguity type	Context
S5	Zipke et al. (2009)	Single sentence	Deep-structure	Flying kites can be exciting.
SP9	Shultz and Pilon (1976)	Sentence pair	Surface-structure	She helped the boy with the hat. She helped the boy put on his hat.
R8	Zipke et al. (2009)	Riddle	Lexical	Why is a school yard larger at recess than at any other time? a. At recess there are more feet in it. b. It isn't.

Table 1: Context type examples

The first type was riddles in which the humour comes from lexical ambiguity. The second type was single, either lexically or structurally ambiguous sentences, including surface-structure and deep-structure ambiguity. The third type of context was sentence pairs in which the first sentence was ambiguous, and the second sentence unambiguously conveyed one of the possible meanings of the first sentence. These sentence pairs also included lexically and structurally ambiguous sentences, incorporating both surface-structure and deep-structure ambiguity.

Kess and Hoppe (1981) define lexical ambiguity as the result "of a word or word sequence having more than one distinct meaning" (p. 30). Surface structure ambiguity "reflects two distinct syntactic groupings of adjacent words in the string … Deep structure ambiguity, on the other hand, reflects different logical relational sets between words or phrases in the sentence." (Kess & Hoppe, 1981, p. 31). The authors point out that in the experiment of Mackay and Bever (1967), the participants "the median perception time for the detection of ambiguities went from lexical to surface structure to deep underlying structure ambiguities", suggesting differences in the difficulty of recognising these three types of ambiguities (Kess & Hoppe, 1981, p. 31).

In this current experiment, there were eight riddles and seven single sentences by Zipke et al. (2009) and 14 sentence pairs by Shultz and Pilon (1973). This resulted in 29 context texts overall. These 29 context texts were used across four task sheets, for a total of 116 tasks per model. The four task sheets (Level 1-Level 4) differed in the style of instructions. Table 2 below contains all instructions based on context type over the four levels.

	Riddles	Single sentences	Sentence pairs
Level 1	Please choose the correct answer to the question and explain your decision.	Please explain the meaning of the sentence.	Please explain the meaning of both sentences.
Level 2	Please choose the correct answer to the riddle and explain your decision. Please also explain why the riddle is funny.	Please explain the possible meanings of the sentence.	Please explain the meaning of both sentences. Please also compare the two sentences.
Level 3	Please explain what makes this a riddle and why it is funny.	Please explain all possible meanings for the sentence.	Please explain all possible meanings for both sentences.
Level 4	Ambiguity makes this riddle funny. Both answers can be considered true. Please explain why. Please also choose which answer is actually true.	The sentence is ambiguous. Please explain all possible meanings for the sentence.	One of the sentences is ambiguous. Which one? Please also explain all possible meanings for both sentences.

Table 2: Instructions for the different levels

As shown, in the case of the riddles, the instructions for Level 2 and Level 3 experimented with ways to encourage the models to analyse the riddles, then contained a direct mention of ambiguity on Level 4. In the case of the single sentences, the instructions contained no reference to ambiguity on Level 1, a hint at ambiguity on Level 2, a more explicit hint on Level 3, and directly referenced ambiguity in Level 4. For the sentence pairs, a combination of encouraging analysis and a direct mention of ambiguity on Level 4 was used. Model responses were elicited via API calls with default parameter settings, which OpenAI do not specify in the available documentation.

3. Results and discussion

The answers of the models were assessed on a binary scale. On Level 1 to Level 3, the assessment was either “Noticed ambiguity” or “Didn’t notice ambiguity”. As on Level 4, ambiguity is addressed in the instruction, and answers were assessed as either “Correct interpretation” or “Incorrect interpretation”. Figure 1 below shows the percentage of correct answers by models and levels:

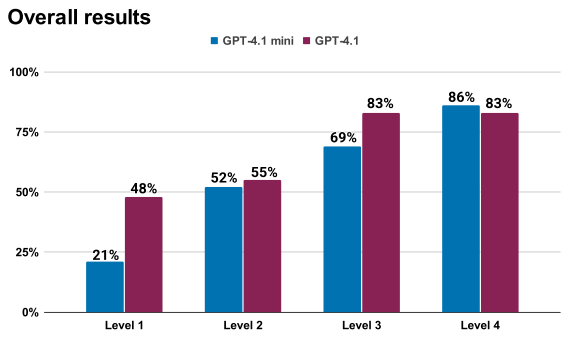

Figure 1: Percentage of correctly solved tasks

As the results show, without at least a hint at ambiguity or encouragement to analyse, the models tended to consider only one possible meaning. With the introduction of referencing the possibility of multiple meanings, both models’ results improved. Interestingly, the results of GPT-4.1 did not increase when the ambiguity was explicitly mentioned at Level 4 compared to Level 3, whereas the results of GPT-4.1 mini improved with every level. While GPT4.1 mini achieved a low score on Level 1, it surpassed GPT-4.1 when ambiguity was explicitly mentioned. Nonetheless, both models showcased significant improvement through more direct prompting.

The two models’ performance varied significantly depending on the type of ambiguity. Figure 2 shows the results for tasks utilising lexical ambiguity:

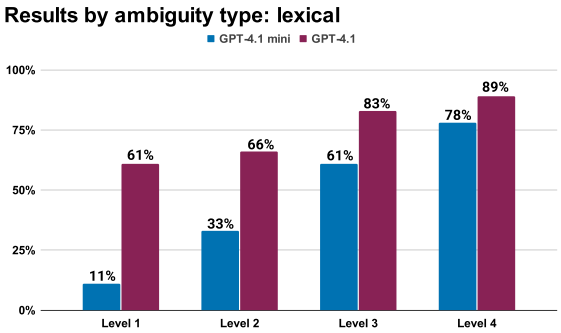

Figure 2: Results of lexical ambiguity tasks

GPT-4.1 consistently outdid GPT-4.1 mini when the task was lexical ambiguity detection. On Level 1, when the instruction does not reference possible double meanings in any way, GPT-4.1 correctly solved 61% of tasks, while GPT-4.1 mini scored only 11%. GPT-4.1 mini showed significant improvement throughout the levels, and GPT-4.1 also continued to steadily improve after the strong start. Nonetheless, as Figure 2 shows, the direct mention of ambiguity did not result in a perfect score for neither of the models.

While the tasks utilising lexical ambiguity seemingly show that the larger model, GPT-4.1, is superior, the results related to structural ambiguity provides a new perspective.

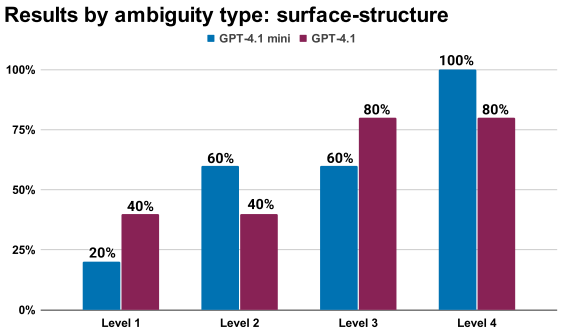

Figure 3: Results of surface-structure ambiguity tasks

As Figure 3 displays, GPT-4.1 mini surpassed GPT-4.1 on Levels 2 and 4. When there was no indication of possible ambiguity, GPT-4.1 outscored GPT-4.1 mini, but on Level 4, GPT-4.1 achieved a perfect score, while GPT-4.1 solved 80% of the tasks correctly. Levels 2 and 3 also highlight an interesting difference between the way instructions affect the models’ performance. GPT-4.1 improved. While GPT-4.1 mini scored better when the possibility of double meanings was offered, the difference between “the possible meanings” and “all possible meanings” did not result in a better performance. In contrast, for GPT-4.1, the difference in instructions between Level 2 and 3 mattered, but the direct mention of ambiguity did not between Levels 3 and 4.

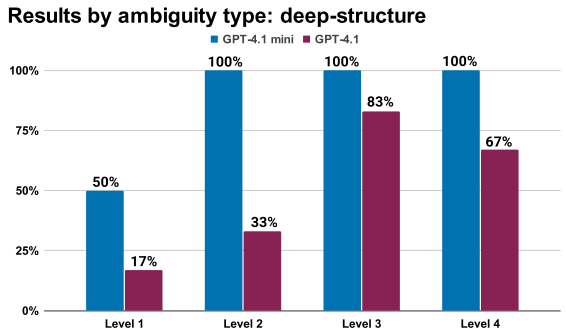

Figure 4: Results of deep-structure ambiguity tasks

Surprisingly, GPT-4.1 mini significantly outperformed GPT-4.1 in deep-structure ambiguity detection. Figure 4 shows that GPT-4.1 mini scored 50% percent on Level 1, and solved all tasks correctly on Levels 2-4, whereas GPT-4.1 scored only 17% percent on Level 1, and its performance peaked on Level 3, with a drop on Level 4.

These results indicate that while GPT-4.1 is better in detecting lexical ambiguity, GPT-4.1 mini surpasses the other model in structural ambiguity detection. This strongly suggest that a larger model size does not necessarily lead to better performance.

As expected, whenever the possibility of ambiguity was implied in the instructions, both models offered various options as facts in all cases, even when they failed to identify all meanings. One such example is related to task SP4 shown in Table 3 below. Neither of the two models considered "tank" as a vehicle but suggested that the gas was put into a gas cannister on various levels. On Level 4, GPT-4.1 mini even identified the second sentence, "He put some gas in the car" as the ambiguous one, and stated that the first sentence, "He put some gas in the tank" is not ambiguous.

Context ID	Original study	Task type	Ambiguity type	Context	Note
SP4	Shultz and Pilon (1976)	Sentence pair	Lexical	He put some gas in the tank. He put some gas in the car.	Both models failed all tasks using this context text. → Neither model considered "tank" as a vehicle, only as "gas tank."
SP11	Shultz and Pilon (1976)	Sentence pair	Deep-structure	It is really quite wonderful to see. It is really a wonderful sight.	Both models failed all tasks using this context text. → Neither model considered the meaning connected to a (lack of) visual impairment.

Table 3: Interesting examples from the results

The examined models' apparent difficulty to detect semantic ambiguity unprompted carries implications relevant to the study of neology. Neology often begins in an ambiguous zone, where a word or phrase is reused in a novel context, and this new use is initially ambiguous or inferable. Then, repeated contextual anchoring stabilises a new meaning, and ambiguity either persists (resulting in polysemy) or resolves (resulting in specialisation). Therefore, ambiguity could be seen as the transitional state of neology before the new meaning reaches high-enough frequency of use (Bybee, 2006).

Metaphorical use of a word or phrase can also result in neologisms (Bowdle & Gentner, 2005). Such examples are "virus" or "cloud" that acquired new, abstract meanings in computing, referencing the base concepts through similarity, resulting in lexically ambiguous words. Early uses were ambiguous and listeners relied on pragmatic inference in uses such as "store your files in the cloud" The sentence "The virus is spreading." remains ambiguous without additional context despite the new meaning having been lexicalised.

Future experiments testing LLMs' semantic ambiguity detection abilities could utilise neologisms, and the results could reveal frequency distribution between base concepts and target concepts. Additionally, LLMs could be tested on their ability to comprehend neologisms

in semantically ambiguous contexts. A possible prompt for an experiment/pilot study could be the following: “In the sentence ‘They left their data in the cloud,’ list all plausible interpretations and rank them by likelihood for usage in 2005, 2010, and 2020.” The answers could be compared to data from tools such as Google’s Books Ngram Viewer.

4. Conclusion

The quantitative evaluation of LLMs presents an ongoing challenge. To contribute to these efforts, this study examined OpenAI’s GPT-4.1 and GPT-4.1 mini models in terms of semantic ambiguity detection and comprehension. GPT-4.1 mini linearly improved with more direct instructions, while GPT-4.1 reached the same percentage of correct solutions on Level 3 and Level 4. The results show that without direct instructions even advanced models, such as GPT-4.1 and GPT-4.1 mini, tend to consider only one possible meaning, most likely based on word or phrase frequency. Additionally, this experiment revealed that GPT-4.1 scores better on tasks utilising lexical ambiguity, while GPT-4.1 mini outperformed the larger model in detecting structural ambiguity, implying that model size is not directly linked to performance. The findings of this experiment could be utilised in neology research, as neologies are oftentimes ambiguous, especially after emergence.

5. Limitations

The results of this study must be considered in the light of its many limitations. Nevertheless, while the experiment was small-scale, the results show that a larger-scale experiment would be beneficial.

Future research plans include building a larger task set with a more balanced inclusion of ambiguity types, as in the current task sheet lexical ambiguity is over-represented. A similar pilot study could be carried out on sentences including recently coined ambiguous neologisms. Furthermore, incorporating a human baseline (native and non-native speakers with varying levels of verified language proficiency) would allow direct comparison between human and LLM abilities in the semantic ambiguity detection domain. To create more reliable assessments, incorporating multiple runs and averaging the results as well as a more nuanced assessment scale are necessary for future experiments. After refining the experimental design, the goal is to test non-commercial as well as other commercially available models.

6. Bibliographical References

Bowdle, B. F., & Gentner, D. (2005). The Career of Metaphor. *Psychological Review*, *112*(1), 193–216. https://doi.org/10.1037/0033-295x.112.1.193

Bybee, J. L. (2006). From Usage to Grammar: The Mind’s Response to Repetition. *Language*, *82*(4), 711–733. https://doi.org/10.1353/lan.2006.0186

Jayaweera, C., & Dorr, B. J. (2025). From Disagreement to Understanding: The Case for Ambiguity Detection in NLI. *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, 37–46. https://doi.org/10.18653/v1/2025.nlperspectives-1.4

Kess, J. F., & Hoppe, R. A. (1981). *Ambiguity in Psycholinguistics* (H. Parret & J. Verschueren, Eds.; pp. 1–123). John Benjamins Publishing.

Miaschi, A., Dell’Orletta, F., & Venturi, G. (2024). Evaluating Large Language Models via Linguistic Profiling. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2835–2848. https://doi.org/10.18653/v1/2024.emnlp-main.166

Roehr-Brackin, K. (2025). Measuring children’s metalinguistic awareness. *Language Teaching*, *58*(1), 27–43. https://doi.org/10.1017/s0261444824000016

Seo, J., Choi, D., Kim, T., Cha, W. C., Kim, M., Yoo, H., Oh, N., Yi, Y., Lee, K. H., & Choi, E. (2024). Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study. *Journal of Medical Internet Research*, *26*, e58329. https://doi.org/10.2196/58329

Shultz, T. R., & Pilon, R. (1973). Development of the Ability to Detect Linguistic Ambiguity. *Child Development*, *44*(4), 728. https://doi.org/10.2307/1127716

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *Npj Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01258-7

Wolters, A., Arz Von Straussenburg, A., & Riehle, D. (2024). *Evaluation Framework for Large Language Model-based Evaluation Framework for Large Language Model-based Conversational Agents Conversational Agents.* https://aisel.aisnet.org/pacis2024/track01_aibussoc/track01_aibussoc/14

Zipke, M., Ehri, L. C., & Cairns, H. S. (2009). Using Semantic Ambiguity Instruction to Improve Third Graders’ Metalinguistic Awareness and Reading Comprehension: An

Experimental Study. *Reading Research Quarterly*, *44*(3), 300–321. https://doi.org/10.1598/rrq.44.3.4

7. Language Resource References

OpenAI, Inc. (n.d.). Models. Platform.openai.com; OpenAI, Inc. Retrieved October 8, 2025, from https://platform.openai.com/docs/models/

Google. (n.d.). *Books Ngram Viewer*. Google.com. https://books.google.com/ngrams/