

Using a Next-Sentence Prediction Task to Test Text Comprehension by Humans and LLMs

Madelein Villegas^{1,*}, Nishitha Uppuluri^{2,*} and Jack Mostow³

¹ De La Salle University, Manila, Philippines

² University of Illinois, Urbana-Champaign, Scarsdale, NY, United States of America

³ Carnegie Mellon University, Pittsburgh, PA, United States of America

Abstract

This paper proposes an automated approach to test text comprehension in both humans and machines. More specifically, it tests inter-sentential processing, which interprets a sentence in the context that precedes it in order to infer its full meaning. Our test uses the Next-Sentence Prediction (NSP) task: namely, given a context, distinguish the subsequent sentence from a distractor sentence. Performance on this task reflects the difficulty of the question and the accuracy of its answer. This task serves multiple purposes:

1. Exercise and test children's inter-sentential processing.
2. Measure the performance of automated text generators such as ChatGPT.
3. Estimate question difficulty.
4. Shed light on comprehension by adding "and why?"

We evaluate this approach on a sample of children's stories by measuring ChatGPT's NSP performance on four types of distractor sentences. The results shed light on the capabilities and limitations of ChatGPT in inter-sentential processing, providing insights for text comprehension assessment.

Keywords

Next-sentence prediction, inter-sentential processing, comprehension test, ChatGPT

1. Introduction

Text comprehension is a crucial skill in knowledge acquisition and language development. Within text comprehension, inter-sentential processing refers to integrating the meaning of a sentence with the meaning of the preceding text [1]. Inter-sentential processing enables readers to establish coherence and make inferences in order to extract the meaning of a text [2, 3].

One method to test inter-sentential processing is the Next-Sentence Prediction (NSP) task [4]. In an NSP task, the reader is given a partial text (the "context") and asked to predict the next sentence. This task tests the ability to integrate information across sentences.

Our original motivation for the NSP task was to add comprehension questions to RoboTutor [5,6], a \$1M Finalist in the Global Learning XPRIZE Competition to develop a tablet tutor to teach basic literacy and numeracy to children with little or no access to schools. RoboTutor already had two forms of automatically generated comprehension questions. A multiple-choice cloze question "Which word comes next?" tests which word fits the syntax, semantics, and context of a sentence. "Which picture shows what you just read?" tests the situation model inferred by the reader from the preceding text [6]. The purpose of the NSP questions is to exercise and test the reader's inter-sentential processing.

Next-sentence prediction questions can be easy to generate but hard to answer. For example, Project LISTEN's Reading Tutor [7] inserted multiple-choice NSP questions at random points in children's

*Work done as members of the RoboTutor team

Proceedings Name, Month XX-XX, YYYY, City, Country

EMAIL: madelein_o_villegas@dlsu.edu.ph (A. 1); nu4@illinois.edu (A. 2); jmostow@andrew.cmu.edu (A. 3)



© 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

stories simply by presenting the next 3 sentences in scrambled order and asking “Which will come next?”. Children’s performance on this task averaged 41%, reliably above chance (33%), but only by a bit [8]. Besides reducing the number of choices from 3 to 2, how can we select next-sentence prediction questions that are easier to answer?

In recent years, Large Language Models (LLMs) have shown advances in natural language processing tasks. Applications using LLMs, such as ChatGPT, possess the ability to generate human-like text and exhibit an impressive understanding of language [9, 10]. This paper assesses ChatGPT’s performance on NSP questions that test inter-sentential processing. We also explore the patterns in ChatGPT’s answers to gain insights into its natural language understanding capabilities and limitations in answering text comprehension questions.

The ability to leverage Generative AI opens up new possibilities for personalized instruction and materials to meet students’ needs [11, 12]. Generative AI can also help teachers improve their effectiveness in teaching by providing targeted feedback and supporting activities that facilitate students’ learning [13]. In contrast, here we use generative AI to test text comprehension as a proxy for human comprehension.

2. Methodology

We now describe our dataset of NSP questions and how we evaluated them.

2.1. Dataset

We derived our dataset from a corpus of 60 English-language children’s stories selected from the African Storybook Project [14] to include in RoboTutor. The initial step involved selecting appropriate context sentences to introduce next-sentence prediction questions. The second author selected context sentences conducive to NSP questions within each story, e.g. neither too long nor too short, focusing on sentences in the middle of the story in order to provide sufficient context to help predict the next sentence, and avoiding those at the beginning where there is too little context, or near the end where there are too few candidates for distractor sentences that have not already appeared. We manually selected 37 context sentences from 34 children’s stories to create the resulting dataset comprised of questions along with 5 sentences chosen automatically from the 10 sentences following each context:

- the “correct” sentence that comes immediately after the context,
- the “nearby” distractor, namely the sentence that follows the correct sentence,
- a “random” sentence chosen randomly from the next 8
- the “easiest” distractor of the remaining 7 sentences, namely the sentence that is the least coherent with the context and hence the easiest to recognize as not fitting,
- and the “hardest” of them, namely the sentence that is the most coherent with the context and hence the hardest to recognize as not fitting.

To determine the “easiest” and “hardest” distractors, we computed coherence scores using the BERT model as described in [4].

2.2. Evaluation

In our study, we evaluate the performance of an application that uses an LLM, specifically ChatGPT, in the context of next-sentence prediction questions. We tried two different prompts: “Which sentence fits better after the context?” and “Which is more coherent?”. To illustrate, consider the following context: “I love soccer so much, what about you? Maliki is teaching us soccer, are you interested too? Let us go to the field.”

The first prompt, "Which sentence fits better after the context?" presents ChatGPT with two candidate successor sentences: "Maliki is the best player!" and "After the game, coach told us that she has a surprise for us." ChatGPT is tasked with determining which candidate sentence fits better after the given context. In the second prompt, "Which is more coherent?" We present ChatGPT with two concatenated texts:

1. "I love soccer so much, what about you? Maliki is teaching us soccer, are you interested too? Let us go to the field. Maliki is the best player!"
2. "I love soccer so much, what about you? Maliki is teaching us soccer, are you interested too? Let us go to the field. After the game, coach told us that she has a surprise for us."

We expect ChatGPT to determine which candidate sentence is more coherent with the given context. We analyzed the model's responses and categorized them into four distinct cases: selecting the correct answer, selecting both candidate sentences as correct, selecting neither candidate as correct, and selecting the wrong distractor sentence. To gain further insights into the model's responses, we added the probe "and why?" after every NSP question. This modification aimed to elicit more detailed explanations from ChatGPT, in order to shed light on the reasoning behind its selected answers.

3. Results and Discussion

ChatGPT's answers to "Which sentence fits better?"

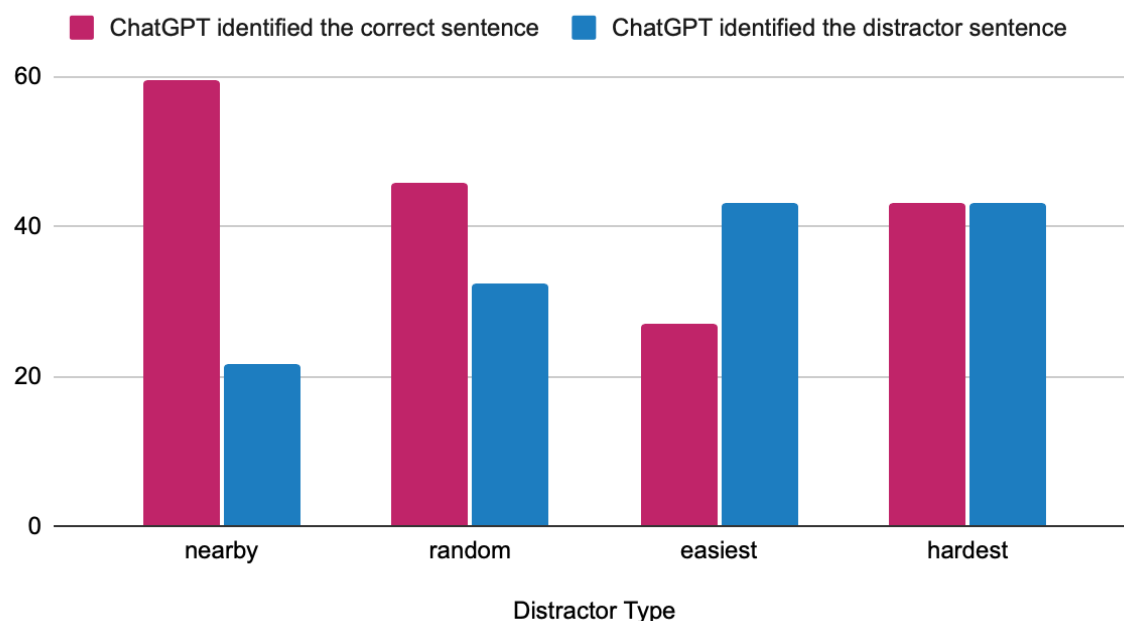


Figure 1: Responses of ChatGPT to the prompt "Which sentence fits better?"

ChatGPT's answers to "Which is more coherent?"

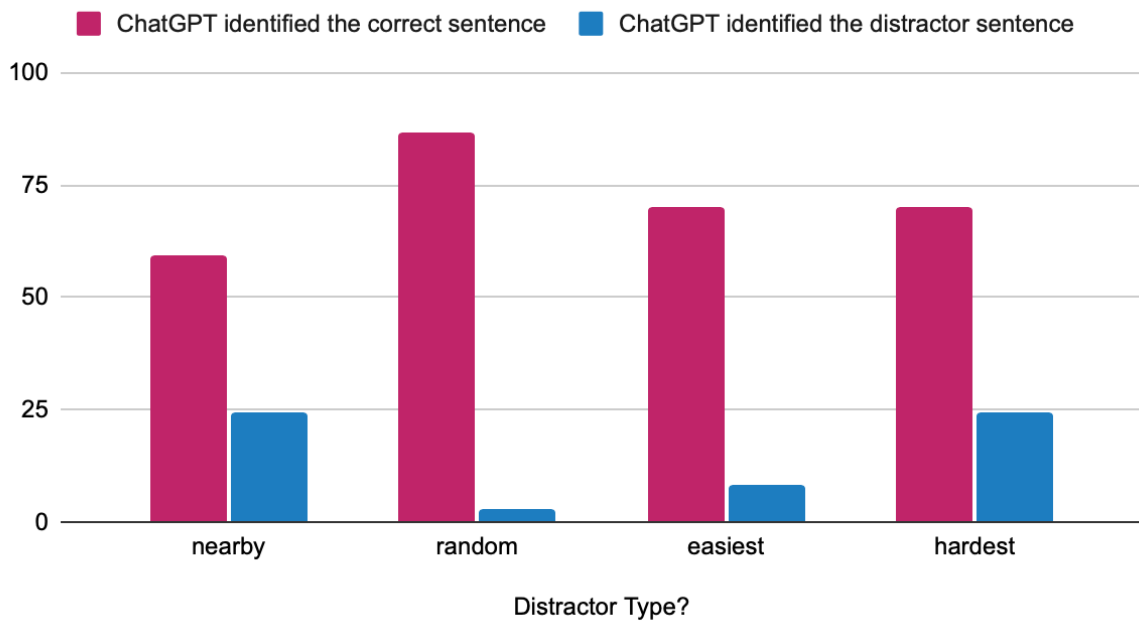


Figure 2: Responses of ChatGPT to the prompt “Which is more coherent?”

3.1. Which sentence fits better?

Our evaluation revealed that ChatGPT demonstrated a weaker performance compared to the other prompt, “Which is more coherent?” in selecting the most suitable sentence given the context. As seen in Figure 1, ChatGPT performs best comparing the “nearby” distractor and the correct sentence, achieving 59.5% correct in classifying the correct sentence as the sentence that follows the context. The “random” and “hardest” distractors perform similarly at 45.9% and 43.2% respectively. Surprisingly, the “easiest” distractor type performed the worst with only 27% of the correct sentences classified as correct. This suggests that ChatGPT has difficulties in distinguishing which sentence follows the context. It is important to note that coherence, which was used to determine the “easiest” and “hardest” distractors, is not the sole measure of determining which sentence fits better.

3.2. Which is more coherent?

ChatGPT performed much better when prompted “Which is more coherent?” As seen in Figure 2, the “random” distractor performs the best with 86.5% correct in classifying the correct sentence with the context when asked for the more coherent series of text, followed by the “hardest” then “easiest” at 70.3% and 70.2% respectively. The similar classification results of the “hardest” and “easiest” distractor types may have been caused by some shared linguistic or semantic characteristics that are difficult for ChatGPT to differentiate, but easy for the BERT model to generate a coherence score. These text features could have obscured ChatGPT’s perceived coherence which allowed the “easiest” and “hardest” distractors to perform similarly. Finally, the “nearby” distractor type performed the best when prompted “which sentence fits better?” but performed the worst when prompted “which is more coherent?”.

3.3. Detailed explanations from ChatGPT

We examined explanations provided by ChatGPT answering the question "and why?"

3.3.1. Correct answer

ChatGPT identifies unexpected jumps in the narrative or instances where a clear connection is lacking. In the excerpt in Table 1, the correct sentence, "They arrived at the river where they needed to cross," establishes a smooth transition from the previous context, indicating a logical progression in the story. The distractor sentence introduces a sudden shift in the narrative without providing a clear connection or explanation. This jump disrupts the flow of the story and leaves a gap in understanding for the reader.

3.3.2. Wrong answer

Wrong answers generated by ChatGPT often contain additional information that is irrelevant to the context sentences. However, ChatGPT may still consider these details relevant, as seen in "...flour to make xima." in Table 1, leading to less plausible explanations. Even if xima is not recognized to be a good or healthy food item, ChatGPT faces challenges in determining whether to introduce new elements to the story or repeat and emphasize existing information. As a result, assumptions about plausible events within the story take precedence over sentences that introduce new information, which may lead to incorrect answers.

3.3.3. Classified both as correct

In cases where both sentences can fit the given context but convey different information, ChatGPT identifies the coherence in both sentences when asked "Which text is more coherent?". When presented with two paragraphs of context sentences along with potential next sentences, both options may appear coherent, and it can be challenging for readers to assess which sentence will follow next. In the example in Table 1, both sentences convey the idea of using the sewing machine to make clothes. The correct sentence states the broader use of the sewing machine while the distractor implies a specific context of the sewing machine's use. In such cases, it can be challenging for readers to accurately predict the sentence that will follow next without additional context, similarly, ChatGPT concludes that both sentences exhibit coherence or can be the next sentence.

3.3.4. Classified neither as correct

When neither of the provided sentences is correct, ChatGPT determines the given sentences do not follow a logical flow or the theme of the story and may generate sentences that may be more appropriate such as "Let's move away from the bull and find a safe place." or "We should call for help to deal with the angry bull." suggested in Table 1. Some sentences may serve as introductory statements for a new idea, making it difficult for ChatGPT to determine the next sentence. Additionally, ChatGPT identifies confusing sentences that lack connection or relevance to the context sentences.

Table 1

Excerpts of Detailed Explanations from ChatGPT

| ChatGPT Classification | Explanations with sentences |
|------------------------|---|
| Correct | <p>Context: There was a bean and piece of coal. They were going to another village.</p> <p>Correct Sentence: They arrived at the river where they needed to cross</p> <p>Distractor: They started to argue about crossing the river.</p> <p>Explanation: The correct sentence is more coherent because it establishes a clear setting and objective for the characters. The distractor sentence</p> |

| | |
|-----------------|--|
| | introduces conflict but doesn't provide a resolution or progression of the plot. |
| Wrong | <p>Context: Mom, can you buy candy?" asked Luisa. "We will buy what we need to make good food, healthy food," replied mother. "We will only buy healthy foods?" asked Luisa.</p> <p>Correct Sentence: That's right.</p> <p>Distractor: First we will buy flour to make xima," said mother.</p> <p>Explanation: The distractor sentence is more coherent because it provides a clear context and a problem that needs to be solved. The correct sentence seems to be disconnected from the previous sentence and lacks context...</p> |
| Both Correct | <p>Context: There is a woman called Mrs Rubandama. She lives in Kakuuto village. She has a sewing machine.</p> <p>Correct Sentence: She uses it to make clothes of different fashions.</p> <p>Distractor: She makes casual clothes for outings.</p> <p>Explanation: Both sentences could fit in the context, but the context sentence seems to be a more general statement about Mrs Rubandama's sewing activities, while the distractor sentence is more specific. Therefore, it depends on the context you want to emphasize. If you want to give a general idea about her sewing activities, you can choose the first sentence. If you want to focus on a specific type of clothing that she makes, then the second sentence would be a better fit.</p> |
| Neither Correct | <p>Context: Be careful! Look over there. That bull is angry with you. Run over here! We will play with you.</p> <p>Correct Sentence: I will jump first.</p> <p>Distractor: You jump after me.</p> <p>Explanation: Neither sentence fits well with the given context because they are unrelated to the situation described. Both sentences are related to jumping, while the context is about being careful around an angry bull. It would be more appropriate to use a sentence related to the situation such as "Let's move away from the bull and find a safe place." or "We should call for help to deal with the angry bull."</p> |

4. Conclusion

We now summarize the contributions of this work, its limitations, and directions for future work.

4.1. Contributions

One contribution of this work is the data set of 37 NSP questions from a sample of children's stories described in Section 2.1 which was a result of an automated approach to assess text comprehension and can be used to evaluate inter-sentential processing. We use the NSP task, where given a context, distinguish the subsequent sentence from a distractor sentence. Performance on this task reflects question difficulty and answer accuracy.

The authors assess ChatGPT's performance on the NSP task using four types of distractor sentences: the "correct" sentence that follows the context, the "nearby" sentence after it, a "random" sentence chosen randomly from the next eight, and the "easiest" and "hardest" distractors based on coherence scores computed using the BERT model.

We also evaluate the performance and accuracy of ChatGPT using the NSP questions created, resulting in insights into the model's natural language understanding capabilities and limitations. In the case of "Which sentence fits better after the context?", ChatGPT performed best in classifying the correct sentence as the one that follows the context, achieving 59.5% accuracy. The "random" and "hardest" distractors had similar performances at 45.9% and 43.2%, respectively, while the "easiest" distractors performed the worst with only 27% accuracy. In the case of "Which is more coherent?", ChatGPT demonstrated better performance. The "random" distractor performed the best at 86.5% accuracy in classifying the correct sentence with the context, followed by the "hardest" and "easiest" distractors at 70.3% and 70.2% accuracy, respectively. Interestingly, the "nearby" distractor performed well in the "Which sentence fits better?" prompt but poorly in the "Which is more coherent?" prompt.

Detailed explanations to the probing question "and why" provided insights into ChatGPT's performance in inter-sentential processing. For example, correct answers exhibited sensibility and logical coherence with the story's sequence, while wrong answers often contained irrelevant information.

Overall, this research offers valuable insights into ChatGPT's performance in inter-sentential processing and has implications for personalized instruction and the use of generative AI in improving text comprehension assessment.

4.2. Limitations

It is important to acknowledge certain limitations in this study. The subjective selection of context sentences may introduce biases or inconsistencies in the selection process, the small dataset size and varying answers depending on the prompt limit the generalizability of the findings, and the evaluation lack external validation. External validation may include correlating the performance of ChatGPT and humans.

4.3. Future work

Extend beyond English: In principle, the NSP task itself is language-independent, insofar as any story comprising a sequence of sentences can be augmented to ask "Which of these sentences is next?"

Question form: Instead of the multiple-choice form of the NSP question, we can simply present either the next sentence or a distractor and ask the Yes/No question, "Is this sentence next?" It's not clear which form of question is easier or better for children. Deciding whether one sentence fits better than another may be easier than deciding whether a sentence fits. On the other hand, comparing two sentences imposes a higher memory load for young readers. We would like to user-test both question forms in RoboTutor to see which form is easier in practice.

And why: The "and why?" probes provide a potential window into ChatGPT's inter-sentential processing to illuminate the opaque LLMs underlying it. LLMs are amazingly intelligent but not always correct. Explaining their behavior in human-understandable terms could be invaluable in analyzing when and how they make mistakes.

5. References

- [1] M. W. Kibby, Intersentential Processes in Reading Comprehension, *J. Read. Behav.* 12.4 (1980) 299–312. doi:10.1080/10862968009547383.
- [2] C. Hall, S. Vaughn, M. A. Barnes, A. A. Stewart, C. R. Austin, G. Roberts, The Effects of Inference Instruction on the Reading Comprehension of English Learners With Reading Comprehension Difficulties, *Remedial Spec. Educ.* 41.5 (2019) 259–270. doi:10.1177/0741932518824983.
- [3] F. Cornish, Inter-sentential anaphora and coherence relations in discourse: a perfect match, *Lang. Sci.* 31.5 (2009) 572–592. doi:10.1016/j.langsci.2008.06.002.

- [4] J. Feng, J. Mostow, Towards difficulty controllable selection of next-sentence prediction questions, in: Educational data mining, 2021.
URL: https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_262.pdf.
- [5] A. McReynolds, S. Naderzad, M. Goswami, J. Mostow, 7th ACM Conference on Learning @ Scale (L@S '20), 2020.
- [6] RoboTutor - RoboTutor - Carnegie Mellon University.
URL: <https://www.cmu.edu/scs/robotutor/>.
- [7] J. Mostow, J. Beck, J. Bey, J. Sison, B. Tobin, J. Valeri, Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions, Technol. Instr. Cogn. Learn. 2 (2004). URL: <https://www.ri.cmu.edu/publications/using-automated-questions-to-assess-reading-comprehension-vocabulary-and-effects-of-tutorial-interventions/>.
- [8] J. Beck, J. Mostow, J. Bey, Can automated questions scaffold children's reading comprehension?, in: Proceedings of 7th International Conference on Intelligent Tutoring Systems (ITS '04).
- [9] H. H. Thorp, ChatGPT is fun, but not an author, Science 379.6630 (2023) 313.
doi:10.1126/science.adg7879.
- [10] Y. Shen, L. Heacock, J. Elias, K. D. Hentel, B. Reig, G. Shih, L. Moy, ChatGPT and Other Large Language Models Are Double-edged Swords, Radiology (2023). doi:10.1148/radiol.230163.
- [11] R. Dilmurod, A. Fazliddin, Prospects for the introduction of artificial intelligence technologies in higher education, ACADEMICIA 11.2 (2021) 929–934. doi:10.5958/2249-7137.2021.00468.7.
- [12] J. Su (苏嘉红), W. Yang (杨伟鹏), Unlocking the Power of ChatGPT: A Framework for Applying Generative AI in Education, ECNU Rev. Educ. (2023) 209653112311684.
doi:10.1177/20965311231168423.
- [13] W. M. Lim, A. Gunasekara, J. L. Pallant, J. I. Pallant, E. Pechenkina, Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators, Int. J. Manag. Educ. 21.2 (2023) 100790. doi:10.1016/j.ijme.2023.100790.
- [14] African Storybook. URL: <https://www.africanstorybook.org>.