Using Next Sentence Prediction to Test ChatGPT's Text Comprehension (Student Abstract)

Ojas M Agarwal[1]*, Madelein Villegas[2], Jack Mostow[3]

[1]Vellore Institute of Technology, Chennai, India
[2]Now at Texas A&M University, College Station, TX, United States of America
[3]Carnegie Mellon University, Pittsburgh, PA, United States of America
ojas.magarwal2021@vitstudent.ac.in, madelein.villegas26@gmail.com, mostow@cs.cmu.edu

Abstract

We propose the Next Sentence Prediction (NSP) task as a simple, objective, scalable, automated way to test ChatGPT's text comprehension. Given a context excerpted from a children's story, the task is to distinguish the next story sentence from a later sentence in the story. We analyze how ChatGPT's performance on this task is related to various features of the text, using data from English and Swahili children's stories.

1 Introduction

Our original motivation for the work reported here was to enhance RoboTutor, a $1M Finalist in the Global Learning XPRIZE Competition (XPRIZE 2024), by adding text comprehension questions. RoboTutor is an automated tablet tutor designed to teach basic literacy and numeracy in English and Swahili to children in developing countries who have limited access to effective human instruction (McReynolds et al. 2020; RoboTutor 2024).

Text comprehension questions vary along multiple dimensions. They can be asked at different times (before, while, or after reading), serve different purposes (Mostow 2011), and test different skills (Mostow et al. 2004; Beck, Mostow, and Bey 2004). In this study, a question is generated to ask after the child reads a sentence, before advancing to the next sentence. Its purpose is to test the child's comprehension while reading a story. The skill it exercises is intersentential processing (Kibby 1980).

Inter-sentential processing helps readers interpret a sentence within the context of the preceding text in order to establish coherence and draw inferences (Hall et al. 2020; Cornish 2009).

As a simple, scalable, automated, multi-lingual way to exercise inter-sentential processing, whether human or automated, we propose the Next Sentence Prediction (NSP) task: distinguish the correct next sentence from a distractor sentence. For both children and ChatGPT, NSP assesses intersentential reasoning by testing the ability to infer from the preceding context which of two sentences is likelier to come next. By definition, the correct answer is the next story sentence, which is trivial to label automatically. This automatic scoring is useful both for RoboTutor in grading student responses and for researchers in analyzing ChatGPT's performance.

The NSP question is intended to exercise inter-sentential processing, not assess it with any precision beyond heuristically classifying the sentence chosen by the child as correct (the next sentence in the story) or incorrect (a distractor sentence from later in the story). Immediate feedback on the child's answer offers the motivational value of the challenge to answer correctly, the cognitive benefit of confirming correct performance, and the psychic reward of being right.

Project LISTEN's Reading Tutor used similar tasks, including cloze questions and sentence prediction, to prompt readers to engage actively with text in order to scaffold their comprehension (Beck, Mostow, and Bey 2004).

Testing ChatGPT on NSP serves as a proxy for human performance, offering a low-cost, scalable way to predict human performance without the expense of collecting actual human data. We analyze how various features of the text affect ChatGPT's performance. These features may also influence human comprehension and thus help predict question difficulty and provide clues about human reading difficulties. Our motivation for the NSP task is educational, but others may find this task useful as a way to test ChatGPT's text comprehension for other purposes.

As a brief illustration, below is an example NSP question to present to a child or input to ChatGPT. It consists of a context excerpted from a 40-sentence story entitled "Animals of Uganda":

...
This is an elephant.
The males have white tusks.
They have large ears.
...

followed by a prompt to choose one of two sentences presented in random order:

Which sentence belongs next after the above context, sentence A or sentence B?
A) Elephants have long trunks.
B) What colour is the elephant?

By definition, the correct answer is whichever sentence actually comes next in the story, in this case sentence A, while distractor B is a sentence from later in the story.

*Postal Address: D2 hostel, VIT Chennai, Kelambakkam-Vandalur road, Chennai-600127; Phone: +918436638966

2 Evaluation

We now evaluate ChatGPT's performance on NSP questions automatically generated from 218 English-language children's stories from the African Storybook Project (ASP 2015). Using ChatGPT-4 version 'gpt-4-turbo' (OpenAI 2024), we generated 12,674 NSP questions based on these stories, which vary in length from 5 to 60 sentences and have diverse settings, characters, and plots.

All the questions used the prompt in Section 1. We also tried 3 other prompts, but they did not change ChatGPT's answers.

We measured performance on the NSP task as percentage correct in distinguishing the next sentence from the distractor. We analyzed how performance varied with these 4 admittedly shallow candidate predictors:

- **Story Length:** The stories ranged from 5 to 78 sentences.
- **Context Length:** We varied the context size from 3 to 10 sentences.
- **Distractor Distance:** The distance between the context and the distractor sentence ranged from 2 to 9 sentences. (The context ends at sentence i and the correct answer is sentence $i + 1$, so the distance between them is 1.) Previous work on automated generation of NSP questions (Feng and Mostow 2021) varied this distance to control their difficulty. The first few sentences after the context were harder to distinguish from the correct next sentence, at least using the BERT large language model.
- **Distractor Length:** Distractor sentences ranged from 2 to 48 words.

The number of NSP questions per story ranged from 5 to 489. Context length and distractor distance were independent variables that we controlled, so we can causally attribute performance differences to them. In contrast, we observed rather than controlled story and distractor length, so their relationships to performance are only correlational.

We split the 218 stories (with 12,674 NSP questions) into 80% training data and 20% test data using scikit-learn (Pedregosa et al. 2011). To analyze ChatGPT's performance on the NSP task, we used 6 methods (decision trees, gradient boosting, linear regression, logistic regression, random forest, and XGBoost) to determine based on the 4 candidate predictors if ChatGPT answered a given NSP question correctly. Logistic regression had the highest predictive accuracy (94%) on held-out test data (2,535 NSP questions), significantly above chance level accuracy (50%) on a T-test (p = 0.017), assuming statistical independence of multiple NSP questions with the same context but different distractors.

Accordingly, we further analyzed logistic regression results. We computed the strength of each predictor in the logistic regression as its logit, and did a T-test of whether it differed significantly from zero. 2 of the 4 predictors were significant: distractor distance (logit 0.063, p=0.047) and distractor length (logit -0.057, p=0.001). That is, further distractors are easier to distinguish from the next sentence, and longer distractors are harder. Context length and story length were not significant.

We computed predictive accuracy on the held-out test data for different values of the 2 significant predictors. That is, we computed for each distractor distance and each distractor length how accurately ChatGPT distinguished the correct next sentence from the distractor.

Predictive accuracy was higher for distractors further from the context, presumably because they tended to be less topically related to it. That is, the sentences immediately following the context were likelier to discuss the same topic than subsequent sentences did. Thus in the example NSP question, the context and sentence A are both about an elephant's anatomy, while sentence B is about its general appearance. Predictive accuracy was stronger for shorter distractors, which surprised us because they provide less information.

To test the generality of our approach, we used 1,010 NSP questions generated for a random sample of 15 Swahili stories from the African Storybook Project. We restricted this sample to stories longer than 10 sentences to ensure that distractor distance could vary from 2 to 9 sentences. Accuracy on the held-out test set of 101 Swahili questions was 100%, yet not significantly above chance (p=0.96), presumably due to the smaller data set (1,010 NSP questions for Swahili vs. 12,674 for English).

3 Conclusion

This paper presents the Next Sentence Prediction (NSP) task as a scalable, automated, language-independent way to test text comprehension. Besides generating NSP questions automatically, this approach automatically labels the next story sentence as the correct answer by definition.

ChatGPT correctly answered 88% of 12,674 NSP questions generated from 218 English stories, significantly greater than the 50% probability of answering correctly by chance. ChatGPT correctly answered 79% of 1,010 NSP questions generated from 15 Swahili stories, although not significantly greater than chance (p = 0.96).

We trained 6 statistical models to predict which questions ChatGPT answered correctly. Logistic regression worked best, achieving 94% predictive accuracy on test data for 42 held-out English stories (better than chance, p = 0.017) and 100% for 5 held-out Swahili stories (albeit too few for statistical reliability). Significant predictors of ChatGPT correctly answering an NSP question from a held-out English story were distractor length (longer are harder) and distractor distance (further are easier). Story length and context length were not significant predictors.

Our data was limited to English and Swahili, with too little Swahili to achieve statistical reliability. Our results are based on children's narrative fiction, so they might not generalize to other genres. We assumed statistical independence of NSP questions generated from the same story context.

Future work could relate ChatGPT's performance to deeper features of the text, such as lexical overlap, sentence structure, and semantic coherence. Perhaps ChatGPT's ability to explain its answers could help predict their correctness better. It would be interesting to test ChatGPT against humans, but human performance varies much more than ChatGPT's, so analyzing it would require much more data. Moreover, human data costs much more to collect.

Acknowledgments

We would like to thank the reviewers for their valuable feedback. Their insightful comments and suggestions greatly contributed to improving its content and clarity.

References

ASP. 2015. African Storybook. https://www.africanstorybook.org.

Beck, J. E.; Mostow, J.; and Bey, J. 2004. Can Automated Questions Scaffold Children's Reading Comprehension? In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems (ITS 2004)*, 478–490. Springer.

Cornish, F. 2009. Inter-sentential Anaphora and Coherence Relations in Discourse: A Perfect Match. *Language Sciences*, 31(5): 572–592.

Feng, J.; and Mostow, J. 2021. Towards Difficulty Controllable Selection of Next-Sentence Prediction Questions. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*.

Hall, C.; Vaughn, S.; Barnes, M. A.; Stewart, A. A.; Austin, C. R.; and Roberts, G. 2020. The Effects of Inference Instruction on the Reading Comprehension of English Learners with Reading Comprehension Difficulties. *Remedial and Special Education*, 41(5): 259–270.

Kibby, M. W. 1980. Intersentential Processes in Reading Comprehension. *Journal of Reading Behavior*, 12(4): 299–312.

McReynolds, A. A.; Naderzad, S. P.; Goswami, M.; and Mostow, J. 2020. Toward Learning at Scale in Developing Countries: Lessons from the Global Learning XPRIZE Field Study. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, 175–183.

Mostow, J. 2011. Questions and Answers about Questions and Answers: Lessons from Generating, Scoring, and Analyzing Questions in a Reading Tutor for Children. https://www.cs.cmu.edu/~listen/2011-11-05%20Mostow-QG2011-keynote.pptx. Invited talk at AAAI Symposium on Question Generation, Arlington, VA, USA.

Mostow, J.; Beck, J.; Bey, J.; Cuneo, A.; Sison, J.; Tobin, B.; and Valeri, J. 2004. Using Automated Questions to Assess Reading Comprehension, Vocabulary, and Effects of Tutorial Interventions. *Technology, Instruction, Cognition and Learning*, 2: 97–134.

OpenAI. 2024. ChatGPT-4. https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

RoboTutor. 2024. RoboTutor: An Automated Tutor for Literacy and Numeracy. https://www.cmu.edu/scs/robotutor/.

XPRIZE. 2024. Global Learning XPRIZE. https://www.xprize.org/prizes/global-learning.