GPT-3.5, GPT-4, Bard, and Claude's Performance on the Chinese Reading Comprehension Test

Bor-Chen Kuo[1], Pei-Chen Wu[1] and Chen-Huei Liao[1]

[1] *National Taichung University of Education, No.140, Minsheng Rd., West Dist., Taichung City 403514, Taiwan.*

Abstract

In this study, we explored the performance of advanced Generative AI models—GPT-3.5, GPT-4, Bard, and Claude—in Chinese reading comprehension tasks. Utilizing a fifth-grade Chinese reading comprehension test, which comprised 55 questions, we assessed the performances of these models in comparison with 491 fifth-grade students from Central Taiwan. The results showed that GPT-4 performed the best in the test and using level settings was more effective than not using them. Analysis of the level settings indicated noticeable differences between Level 1 and 2 for GPT and Bard, with less distinct variations observed between Level 2 and 3. In contrast, Claude exhibited minimal variation in results across all levels. The performance of the human students was similar to that of GPT-3.5, but not as that of high as the other models. For future research, we recommend employing a more nuanced design for prompts to better simulate the reading comprehension abilities of students of various ages, thereby further enhancing the educational applications of these models.

Keywords

large language models, reading comprehension, pass rate

1. INTRODUCTION

In recent years, language models have rapidly evolved from early iterations such as BERT, GPT, and GPT-2 to GPT-3, signifying the onset of the era of large-scale language modeling. The GPT-3 model, with its 175 billion parameters, has been trained on a substantial dataset, enabling its application across a broad spectrum of domains without the need for specialized training [1]. However, models designed for specific tasks can yield more precise results. Due to their advanced capabilities, large-scale language models are increasingly utilized in educational settings, helping to generate questions, create text, understand the language, and automated grading [2].

This study aims to evaluate the performance of generative models such as GPT-3.5, GPT-4, Bard, and Claude in Chinese reading comprehension tasks. Its primary objective is to determine if these models can accurately simulate the reading comprehension skills of students at different levels. Furthermore, the study will compare the performance of these generative models with that of human students in similar reading comprehension tasks.

Based on the above research objectives, the research questions of this study are as follows:

RQ1: How does the performance of GPT-3.5, GPT-4, Bard, and Claude vary with and without level settings?

RQ2: What is the performance of GPT-3.5, GPT-4, Bard, and Claude in the Chinese Reading Comprehension Test at different levels?

RQ3: How does the performance of GPT-3.5, GPT-4, Bard, and Claude compare to that of human students in Chinese reading comprehension test?

2. METHODS

In this study, we employed the fifth-grade Chinese reading comprehension test developed by Prof. Chen-Huei Liao's team at National Taichung University of Education [3] as a test tool. This test was used to evaluate the performance of various language models – GPT-3.5, GPT-4, Bard, and Claude – in Chinese reading comprehension. Our goal was to determine how effectively these models simulate

Joint Proceedings of LAK 2024 Workshops, co-located with 14th International Conference on Learning Analytics and Knowledge (LAK 2024), Kyoto, Japan, March 18-22, 2024.

kbc@mail.ntcu.edu.tw (B.C. Kuo); pedropcwu@gmail.com (P. C. Wu); chenhueiliao@gmail.com (C. H. Liao)

0000-0003-1741-2450 (B.C. Kuo) ; 0009-0000-1343-5407 (P. C. Wu)

reading comprehension across different levels and to compare their pass rates with those of human students.

The test consists of 55 questions, characterized by an average difficulty of 0.614, a discrimination of 0.39, and a reliability of 0.899. It includes four question types: word and phrase, sentence, contextual comprehension, and inference, covering six dimensions: phonological processing ability, vocabulary comprehension, sentence comprehension, grammatical comprehension, contextual comprehension, and inferential comprehension. The format is a four-option multiple-choice test.

According to the research objectives, the following tasks will be carried out in this study:

1. T1：Evaluate the effects and performance of GPT, Bard, and Claude in Chinese reading comprehension test with and without level settings.

2. T2：Compare the performance of GPT, Bard, Claude, and human students in Chinese reading comprehension test.

2.1 T1 TEST

The purpose of this test was to address Research Questions 1 and 2 (RQ1 and RQ2), specifically to evaluate the effects and response results of the model both with and without level settings. The aim was to ascertain whether the model could effectively simulate reading comprehension test performance for students at different levels. In this study, the levels were defined to represent various age groups: Level 1 for grades 1 to 3, Level 2 for grades 4 to 6, and Level 3 for grades 7 to 9. The initial test was conducted without a level setting. The same prompt was inputted into all four models, with the prompt set as follows: 'You are now asked to do a reading comprehension test, please solve the question, there are 55 questions in total, and they will be provided in batches.'

We discovered that the model's effectiveness in answering the questions diminished when it was given all 55 questions at once. The slower response speed could be attributed to the challenge of processing a large amount of text simultaneously, which appeared to decrease its parsing ability and increase the error rate in question-solving. Consequently, we decided to present 10-15 questions at a time to the model and then calculated the pass rate by comparing the selected answers with the correct ones.

In the next phase of testing, which included level settings, all four models were given the same prompt, intending to have each model simulate the reading comprehension level of students of different grades. Taking Level 2 as an example, the content of the prompt was: 'You are now a Grade 4 - 6 student, and you are now asked to do a reading comprehension test based on the reading comprehension skills you should have at your current level. There are 55 questions in total, in total, and they will be provided in batches.' This approach was consistent with the previous one. We found that if the model was tasked with answering all 55 questions at once, its effectiveness decreased. The potential lower parsing ability when reading large texts at once could lead to a higher error rate in solving the questions. Moreover, when simulating students of different grades, the results were nearly identical for students in grade 4 and above, making it challenging to distinguish between the reading comprehension abilities of students in different grades. Ultimately, we again opted to provide the model with 10-15 questions at a time, recording the response options and the correct answers to calculate the pass rate.

2.2 T2 TEST

The objective of this test was to address Research Question 3 (RQ3), which aimed to compare the performance of the model with that of human students on a Chinese reading comprehension test. The model's response data were sourced from the T1 TEST. For human students' response data utilized in this study were obtained from Lin [3], which involved the participation of 491 fifth-grade students in Central Taiwan. This assessment was conducted using a paper-based format. After the testing, the students' responses were digitized. The data were then subjected to a detailed analysis using BILOG-MG, culminating in the calculation of the average pass rate among the students, based on the results of this analysis.

3. RESULTS

The results demonstrated that all four models exhibited improved performance with level setting compared to without. GPT-4 emerged as the top performer, followed by Claude, then Bard, and finally GPT-3.5, as illustrated in Table 1.

Table 1
The results of Without/With Level Setting

Model	Level Setting	Pass rate
GPT3.5	N	67.27%
	Y	67.88%
GPT 4	N	85.45%
	Y	87.88%
Bard	N	70.91%
	Y	72.12%
Claude	N	78.18%
	Y	80.61%

Note. With level Setting (Y) indicates the average pass rate of the level.

According to Table 2, when GPT, Bard, and Claude are given the same prompt, the pass rates for GPT and Bard exhibit notable variation at different levels, particularly between Level 1 and Level 2. In contrast, Claude shows negligible variation (only a 1.82% difference between Level 1 and Level 2). During the testing phase, the Claude model indicated that it cannot fully replicate the cognitive and problem-solving abilities of students of a specific age group. However, it can attempt to solve problems by employing basic vocabulary and knowledge suitable for that age group, complemented by relevant assumptions and inferences. The final outcomes align with the initial descriptions provided by the model.

Table2
Pass rates at different levels for different models

Model	Level	Pass rate
GPT-3.5	Level 1	65.45%
	Level 2	69.09%
	Level 3	69.09%
GPT- 4	Level 1	81.82%
	Level 2	90.91%
	Level 3	90.91%
Bard	Level 1	69.09%
	Level 2	72.73%
	Level 3	74.55%
Claude	Level 1	81.82%
	Level 2	80.00%
	Level 3	80.00%

In the final comparison between the model's performance and that of human students, it was found that the pass rate for human students stood at 67.41%, most closely aligning with the performance of GPT-3.5.

4. RESULTS

The results of the study showed that GPT-4 performed the best on the test, with level setting being more effective than without level setting. The analysis of the level setting revealed a more pronounced difference between Level 1 and Level 2 for GPT and Bard, whereas the difference between Level 2 and

Level 3 was less marked. The performance of Claude in Level 1, 2, and 3 was similar. This suggests that Claude was less adept in this capacity. The performance of the human students was similar to that of GPT-3.5, but not as good as the other models.

For future enhancements, in addition to fine-tuning the model, we can consider specifying the reading comprehension abilities expected of students in different age groups when providing the prompt. This strategy could more accurately align the model with the actual thinking and problem-solving patterns of students across various age groups during simulation.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1877-1901. doi:10.48550/arXiv.2005.14165.

[2] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, et al., ChatGPT for good? On opportunities and challenges of large language models for education, Learning and Individual Differences 103 (2023) 102274. doi: 10.1016/j.lindif.2023.102274.

[3] W.C. Lin, Establishment of the computerized adaptive reading comprehension test for fifth grade students in elementary school, Master's thesis, National Taichung University of Education, 2014. URL: https://hdl.handle.net/11296/z2xa8e.