AI Skill Check

An Examination of ChatGPT's Consistency in Linguistics and Mathematics using the SAT

Ma. Elisha Kaye B. Nazario
De La Salle University Integrated School
Manila, Philippines
maria_elisha_nazario@dlsu.edu.ph

Eliza Janel L. Tan
De La Salle University Integrated School
Manila, Philippines
eliza_janel_tan@dlsu.edu.ph

Steven Vincent D. Lim
De La Salle University Integrated School
Manila, Philippines
steven_vincent_lim@dlsu.edu.ph

Keith Alexandre L. Ramos
De La Salle University Integrated School
Manila, Philippines
keith_ramos@dlsu.edu.ph

Shirley B. Chu
De La Salle University
Manila, Philippines
shirley.chu@dlsu.edu.ph

ABSTRACT

In the era of technological boom, the use of artificial intelligence is becoming increasingly intertwined with everyday societal systems. This includes the integration of Open AI's ChatGPT into the way students learn today. As a result, the academic community has continuously assessed the performance of ChatGPT in various domains, especially in linguistics-related and mathematics-related fields. However, there still exists a gap to allow a more diverse understanding of ChatGPT's consistency in answering questions on these domains at a high-school level. Therefore, this study aims to quantify the extent to which ChatGPT may generate inconsistent responses to high-school-level linguistics and mathematics questions when fed standardized SAT questions commonly used to evaluate a high-school student's knowledge and aptitude. By extension, it provides a more detailed analysis of ChatGPT's potential as a learning tool. Through this investigation, it was observed that ChatGPT generally demonstrates greater consistency in linguistics compared to mathematics, with different levels of reliability across distinct SAT subareas.

KEYWORDS

ChatGPT, SATs, consistency, high-school, learning tool

ACM Reference Format:
Ma. Elisha Kaye B. Nazario, Eliza Janel L. Tan, Steven Vincent D. Lim, Keith Alexandre L. Ramos, and Shirley B. Chu. 2024. AI Skill Check: An Examination of ChatGPT's Consistency in Linguistics and Mathematics using the SAT. In *Proceedings of Philippine Computing Science Congress (PCSC2024)*. Laguna, Philippines, 8 pages.

1 INTRODUCTION

1.1 Background of the Study

The rapid development of technology over history has led to the rise of artificial intelligence (AI) that continues to unlock new opportunities in the education sector, benefitting both educators and life-long learners as affirmed by the United Nations Educational, Scientific and Cultural Organization or UNESCO, on Sustainable Development Goal 4, focusing on Quality Education [14]. However, despite the convenience it may bring, the use of AI also prompted concerns and risks regarding its ethical usage, security, and long-term implications for student learning among its users. One such discourse revolves around ChatGPT. Its popularization in usage among students has been evident since its release in November 2022. [8] noted ChatGPT's instant feedback capability prompts excessive reliance on it as a learning tool for Linguistics and Mathematics (subjects often taken at the high-school level). This is despite the risk of unintentional misinformation.

ChatGPT works by drawing from datasets up to September 2021. It then refines its responses using human feedback. This ability of ChatGPT prompted studies like [4] and [13] to analyze ChatGPT's proficiency in test-taking scenarios. However, more scrutiny must still be done on its performance in high school contexts.

This highlights the significance of the College Board's SATs. This standardized test gauges high school students' college readiness, with over a million students taking it annually [1, 9, 12]. As students start utilizing ChatGPT for SAT preparations, understanding its inconsistencies, aside from its accuracy, is important to understand its effectiveness as a learning tool [2, 10].

This study aims to quantify inconsistent responses generated by ChatGPT when presented with high-school-level linguistics and mathematics questions.

1.2 Scope and Limitations

This study evaluates ChatGPT's performance on selected data sets involving linguistics and mathematics SAT practice test questions gathered from both offline and online sources. Only questions with a defined answer, such as multiple-choice linguistics questions, multiple-choice mathematics questions, and free-response mathematics questions, are considered. The analysis of ChatGPT's performance in this study mainly focuses on its responses' consistency, to determine if it can consistently answer high-school level mathematics and linguistics questions correctly or incorrectly. Due to ChatGPT's limitation involving images and geometric shapes, questions involving these are excluded.

This study also only covers understanding the performance of GPT-3.5 as this is the only version that is accessible for free by high school students.

PCSC2024, May 2024, Laguna, Philippines

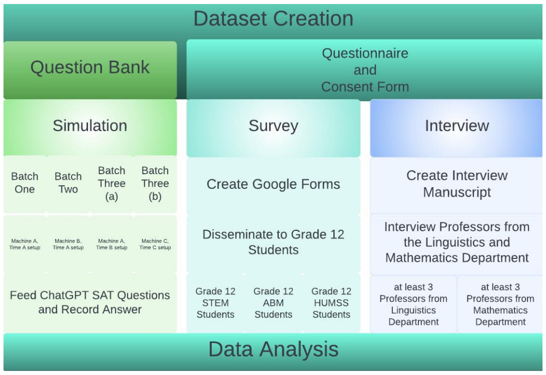

Dataset Creation
Question Bank
Questionnaire and Consent Form
Simulation
Survey
Interview
Batch One
Batch Two
Batch Three (a)
Batch Three (b)
Create Google Forms
Create Interview Manuscript
Machine A, Time A setup
Machine B, Time A setup
Machine A, Time B setup
Machine C, Time C setup
Disseminate to Grade 12 Students
Interview Professors from the Linguistics and Mathematics Department
Feed ChatGPT SAT Questions and Record Answer
Grade 12 STEM Students
Grade 12 ABM Students
Grade 12 HUMSS Students
at least 3 Professors from Linguistics Department
at least 3 Professors from Mathematics Department
Data Analysis

Figure 1: Schematic Representation of Study's Methodology

2 RELATED WORKS

A number of studies have delved into ChatGPT's performance and limitations across various academic disciplines. However, most focus on ChatGPT's accuracies in mathematics and linguistics in the context of post-secondary level. In mathematics, [11] highlights ChatGPT's ability to solve mathematical word problems. [6] explores its proficiency in advanced math, revealing both accurate and partially incorrect answers. On the other hand, [3] and [5] reveal that ChatGPT's linguistics proficiency is comparable to an average student in producing clear, concise responses in academic writing. Both cases show that ChatGPT has the ability to comprehend linguistic and mathematical expressions.

While there is limited research that directly studies the consistency of linguistics and mathematics, [6] has noted that while ChatGPT can indeed demonstrate enhanced language understanding abilities and deductive reasoning ability, it can still make mistakes that violate logical properties and that it can sometimes change its answer if a question is paraphrased.

3 METHODOLOGY

3.1 Research Design

While the complete study utilizes a mixed-method triangulation approach, this paper will focus on the quantitative approach, specifically testing ChatGPT through simulations. The entire procedure has been illustrated in Figure 1.

3.1.1 Dataset Creation. The dataset creation phase involves the compilation of a question bank to store and categorize SAT questions for linguistics and mathematics sets from different online and offline sources and record ChatGPT's response to each question. Questions are categorized and labeled by source and subject.

These datasets are randomly split into three batches of 40 questions for each domain. Each batch is divided equally among subareas of each domain.

In linguistics, questions are sourced from the Reading Test, and Writing and Language Test sections of the SAT. In each batch, ten questions come from each of the following subareas:

(1) Standard English Conventions, which focuses on sentence structure, usage, and punctuation;

Table 1: Example Revisions Made to Linguistics Questions for the Simulation

Original	Input to ChatGPT
A goat ingests the vegetation particular to the meadow in which it grazes, which, along with other environmental **factors such as altitude and weather** shapes the cheese's taste and texture.	A goat ingests the vegetation particular to the meadow in which it grazes, which, along with other environmental factors such as altitude and weather shapes the cheese's taste and texture.
A) NO CHANGE B) factors, such as altitude and weather, C) factors such as, altitude and weather, D) factors, such as altitude and weather	**What improvements can be made to "factors such as altitude and weather"?** A) NO CHANGE B) factors, such as altitude and weather, C) factors such as, altitude and weather, D) factors, such as altitude and weather

(2) Expression of Ideas, which touches upon topic development, organization, and rhetorically effective use of language;
(3) Relevant Words in Context, which focuses on addressing word/phrase meaning in context and rhetorical word choice;
(4) Command of Evidence, which assesses the interpretation and usage of evidence found in passages and informational graphics (e.g. graphs, tables, and charts).

For mathematics, questions encompass two types: multiple choice and response. Ten questions from each of the following subareas in the mathematics section form each batch of dataset:

(1) Heart of Algebra, which involves linear equations and inequalities questions;
(2) Problem Solving & Data Analysis, which tests quantitative reasoning and the interpretation of data (ratio and percentages);
(3) Passport to Advanced Math, which focuses on understanding expression structure, reasoning with more complex equations, and interpreting and building functions;
(4) Additional Topics in Mathematics, which focuses on other questions, including trigonometry and geometry.

Since input to ChatGPT is limited to text only, some questions are modified. Additional instructions are included to replace the underlined words or phrases in the original texts (see Table 1). LaTeXwas used for typesetting mathematical expressions that have complex symbols and structures. A sample is shown in Table 2.

3.1.2 Simulation. The simulation phase involves inputting the questions from the dataset to ChatGPT and recording its responses. Inspired by methodologies in [5, 11, 15], ChatGPT's consistency in a topic is examined by having it answer each question four times under different conditions or on different machines. The setups

Table 2: Example Revisions Made to Mathematics Questions for the Simulation

Original	Input to ChatGPT
The volume of a sphere is given by the formula $V = \frac{4}{3}\pi r^3$ where r is the radius of the sphere. Which of the following gives the radius of the sphere in terms of the volume of the sphere?	The volume of a sphere is given by the formula **V=\frac{4}{3}\pi r^3** where *r* is the radius of the sphere. Which of the following gives the radius of the sphere in terms of the volume of the sphere?
A) $\frac{4\pi}{3V}$	A) **\frac{4\pi}{3V}**
B) $\frac{3V}{4\pi}$	B) **\frac{3V}{4\pi}**
C) $\sqrt[3]{\frac{4\pi}{3V}}$	C) **\sqrt[3]{\frac{4\pi}{3V}}**
D) $\sqrt[3]{\frac{3V}{4\pi}}$	D) **\sqrt[3]{\frac{3V}{4\pi}}**

Table 3: Independent Variables per Setup for Comparison

	Setup A (CONTROL) Time	Independent Location
Setup B		✓
Setup C	✓	
Setup D	✓	✓

include a Control Group (**Machine A, Time A**) for benchmarking, Machine Variation (**Machine B, Time A**) to assess machine-dependent responses, Temporal Variation (**Machine A, Time B**) for understanding temporal stability, and Location and Time Variation (**Machine C, Time C**) to explore external factors' influence on performance. See Table 3 for a comparison of independent variables per setup.

During the simulation, ChatGPT was manually fed SAT questions. However, each batch utilized additional prompts to test the effects of the prompts on ChatGPT's consistency. A "Give answer only" prompt was added to some questions in Batch 1, not included in Batch 2, and was added to all questions in Batch 3. Moreover, there are instances when ChatGPT changes its answer halfway through its explanation, i.e. its answer at the beginning, is different from its answer at the end of its explanation or its solution. In such cases, ChatGPT's initial response is recorded.

3.2 Data Analysis

In order to compare and examine ChatGPT's consistency in answering questions correctly for each SAT subarea, a manual counting of ChatGPT's answers is performed. The number of times ChatGPT answered correctly in one setup, two setups, three setups, and all setups, or none of the setups, is also noted.

Additionally, to compare the consistency between linguistics and mathematics, the standard deviation of all setups per batch is calculated along with the total mean standard deviation.

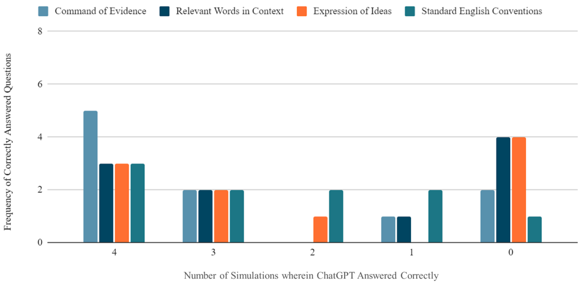


Figure 2: Batch 1 Results

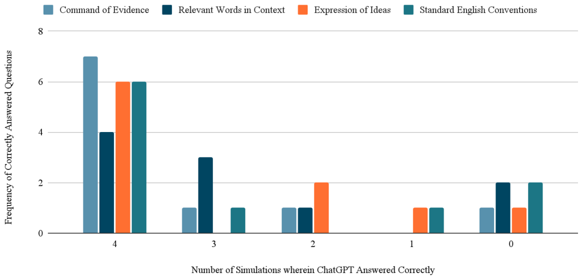


Figure 3: Batch 2 Results

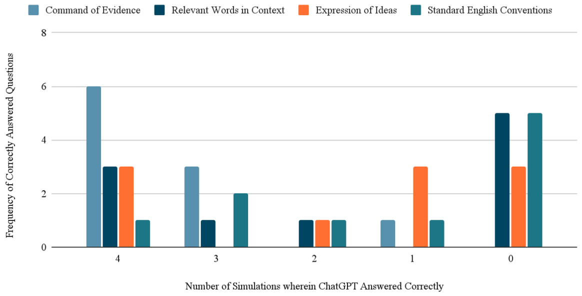


Figure 4: Batch 3a Results (with prompt)

4 RESULTS AND DISCUSSION

4.1 Linguistics

This chapter summarizes ChatGPT's consistency in each of the SAT subareas for each batch and its overall consistency in each domain through the calculated mean standard deviation. The values gathered were based on ChatGPT's raw accuracy score when it was fed SAT questions for each setup and each batch.

Consistency is defined as ChatGPT's tendency to generate the same response when given the same prompt. Hence, the terms *consistently correctly* and *consistently incorrectly* shall be used in instances where ChatGPT was able to get the answer to a question correctly or incorrectly for all four setups of each batch.

Figure 2 shows that in Batch 1, ChatGPT is *consistently correct* in COMMAND OF EVIDENCE, being able to answer five questions

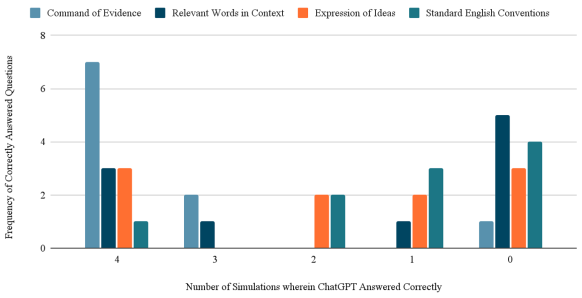


Figure 5: Batch 3b Results (without prompt)

Figure 6: ChatGPT's Consistency in the SAT Linguistics Subareas

correctly in all four machines. The data also shows that ChatGPT is *consistently incorrect* in the subareas of Relevant Words in Context and Expression of Ideas. This indicates that during the simulation of the first batch of questions, the software for all machines was more capable of noting details and answering questions about provided texts. However, it is relatively weak in determining the meanings of words and effectively expressing complete thoughts. Additionally, results show that in the Standard English Conventions subarea, ChatGPT answered more questions *consistently correctly* than incorrectly in all machines.

Figure 3 shows that in Batch 2, ChatGPT answered questions *consistently correctly* more than incorrectly. This also shows ChatGPT's improvement in all of the four subareas. Based on the Batch 2 results, ChatGPT answers most *consistently correctly* questions under Command of Evidence. In this simulation, it answered seven questions correctly in all machines compared to five questions in the previous simulation. There is also an increase in correctly answered questions compared to the previous batch, which shows ChatGPT's strength in answering questions about details in texts. Following Command of Evidence, ChatGPT consistently answers six questions correctly in Expression of Ideas and Standard English Conventions. This improvement from the previous batch's results shows that ChatGPT is able to effectively utilize the English language to convey thoughts and identify grammatical errors in texts. Lastly, in all simulations, ChatGPT answers with the fewest questions correct in the Relevant Words in Context subarea.

Figure 4 shows Batch 3a results. The "Give the answer only" prompt is included in all questions in this simulation. With the additional prompt, ChatGPT's results changed significantly compared to the previous two batches. While ChatGPT maintains the highest consistency of correctly answered questions in the subarea of Command of Evidence, results in the other three subareas begin to show its weaknesses. Firstly, it answered more questions consistently incorrectly in the subareas of Relevant Words in Context and Standard English Conventions. Its worst performance is in the latter, wherein only one question in all machines is answered correctly. The batch's results seem to show that restricting ChatGPT's response to a selection between four options without giving it the freedom to explain its answers affects its answers significantly. This is especially evident in the subareas of Relevant Words in Context and Standard English Conventions, wherein any form of explanation would be needed to understand the thought process that led ChatGPT to determine the definitions of words or the errors in sentence structure. In the Expression of Ideas subarea, ChatGPT neither consistently answers more questions correctly nor incorrectly.

When testing the same set of questions without the "Give answer only" prompt, there is no significant difference with regard to ChatGPT's consistency (Figure 5). ChatGPT only consistently answers one additional question correctly across all batches under Command of Evidence and is able to answer more questions correctly in some batches under the Standard English Conventions subarea. This suggests that in linguistics, either prompt does not significantly affect consistency or a more appropriate prompt must be used.

4.1.1 Consistency for each Subarea. After analyzing the consistency of ChatGPT's answers for the linguistics simulations, the following can be observed:

(1) ChatGPT answers most *consistently correctly* in questions under the Command of Evidence subarea. This is seen in all four simulations. This demonstrates ChatGPT's capabilities in comprehending, noting details from, and answering questions about provided texts, a competency observed by other studies that tested its performance in answering reading comprehension questions, such as [5].

(2) ChatGPT answers most *consistently incorrectly* in the Relevant Words in Context subarea in all simulations. This indicates ChatGPT struggled in discerning the definitions of words used in sentences. This observation is quite ironic, as researchers such as [7] have noted its remarkable ability to generate accurate definitions for various words similar to the Collins Birmingham University International Language Database (COBUILD). Considering that some of the questions were offline-sourced and that limitations of text inputs meant that the words asked could not be highlighted in the questions themselves, there are still some factors that contribute to ChatGPT's weakness in answering the questions under this subarea. While prompt engineering is outside the scope of this paper, other prompts may be identified to increase ChatGPT's consistency.

(3) Adding the "Give the answer only" prompt to the questions in Batch 3 does not make a difference. This only implies that ChatGPT's consistency is not affected by this prompt. It does not necessarily mean that ChatGPT performs better without it, as that is beyond the scope of this paper.

4.2 Mathematics

It can be seen in Figure 7 that ChatGPT does not have a pattern in terms of consistency in any of the mathematics subareas. Notably, it was not able to get any questions correct across all machines in the Additional Topics in Mathematics subarea. Moreover, it could only answer five questions correctly two times in the Heart of Algebra subarea. ChatGPT performs the best in Passport to Advanced Math, not being *consistently incorrect* in any question

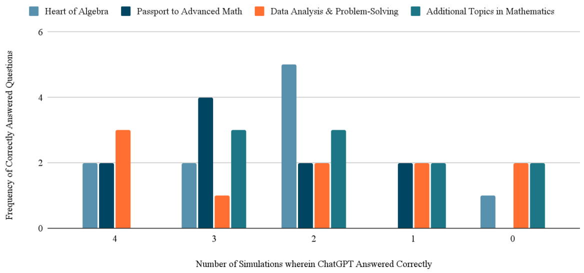


Figure 7: Batch 1 Results

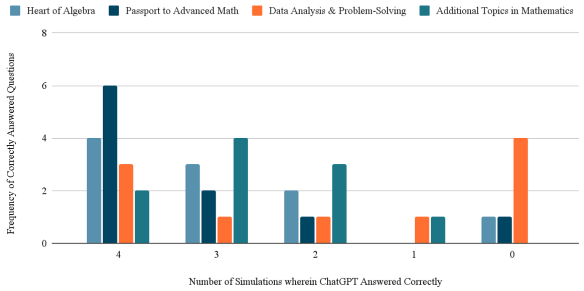


Figure 8: Batch 2 Results

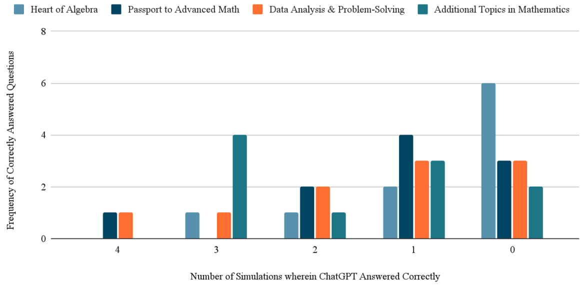


Figure 9: Batch 3a Results (with prompt)

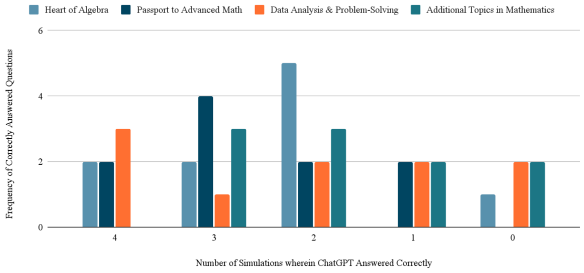


Figure 10: Batch 3b Results (without prompt)

Figure 11: ChatGPT's Consistency in the SAT Mathematics Subareas

and getting four questions correct three times, albeit only being able to answer two questions correctly across all batches.

Figure 8 shows that ChatGPT exhibits significant improvement in Batch 2 in terms of consistency. It has the largest improvement in Data Analysis and Problem Solving and Additional Topics in Mathematics. This may mean that ChatGPT may have been updated between the Batch 1 and Batch 2 simulations.

In contrast, the results of Batch 3a in Figure 9, which utilizes the "Give answer only" prompt, show a drastic performance decline in all subareas. Notably, ChatGPT answered *consistently incorrectly* six questions in Heart of Algebra subarea. In Passport to Advanced Math, it only answered four questions correctly once. Moreover, it got no questions correct in all batches in Heart of Algebra and Additional Topics in Mathematics. This illustrates that using the "Give answer only", which prevents ChatGPT from generating solutions, increases its tendency to be *consistently incorrect.*

Without the "Give answer only" prompt, ChatGPT's consistency returns to the similar trend observed in Batches 1 and 2. Figure 10 shows that ChatGPT answered six questions *consistently correctly* in all four machines in the Passport to Advanced Math subarea. Results in this simulation show a decline in performance compared to Batch 2.

After analyzing the consistency of ChatGPT's answers for the mathematics simulations, the following can be observed:

(1) ChatGPT does not exhibit any clear trend regarding consistency and inconsistency in the mathematics subareas. For instance, Passport to Advanced Math could be the most consistently correct subarea for Batch 1, but it is the least consistently correct by Batch 2. This supports the findings of [6], which posits that ChatGPT struggles to do accurate calculations consistently and falls short compared to chatbots specifically trained to do mathematics.

(2) Succeeding batches show overall improvement. Batch 2 and 3b exhibit greater consistency in answering more questions correct as compared to Batch 1. This could imply that ChatGPT possibly improved over time due to the inputs it receives.

(3) ChatGPT performs significantly worse if it is prompted only to give an answer without any solutions using the "Give answer only" prompt. It may be inferred that ChatGPT has a higher chance of arriving at a correct answer if it is given the liberty to generate a thought process or solution. It is worth noting that its overall performance may change if a different prompt is used to ask ChatGPT not to show any solution.

(4) ChatGPT may not be suitable for mathematical contexts. ChatGPT is a text-oriented LLM chatbot that follows patterns from its online database. The nature of math problems changes depending on the wording and values provided in the question.

4.3 Comparison of ChatGPT's Consistency between Linguistics and Mathematics

To determine in which domain—linguistics or mathematics, ChatGPT exhibits more consistency, the standard deviation of ChatGPT's

Table 4: Standard Deviation of ChatGPT's Achieved Linguistics Mathematics Scores per Batch

	Linguistics	Mathematics
Batch 1	1.25	10.05
Batch 2	2.72	3.95
Batch 3a	5.12	9.60
Batch 3b	5.41	2.80
Mean Standard Deviation	**3.63**	**6.60**

Table 5: ChatGPT's Answer to a Mathematics Question with and without prompt

Question number: **1**				
Expected answer: **D**				
	CS[*]	TV[@]	MV[#]	MTV[°]
Batch 3a[a]	C	B	A	B
Batch 3b[b]	D	D	D	D

[*] Control setup
[@] Temporal variation
[#] Machine variation
[°] Temporal and Machine variation
[a] with prompt
[b] no prompt, may show solution

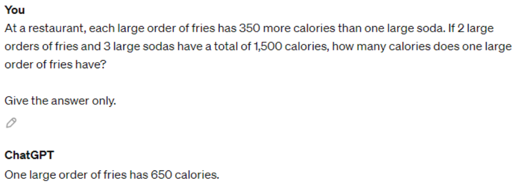
score in each batch is computed. Table 4 lists the standard deviation of all setups for all batches. A lower standard deviation value indicates higher consistency.

Data shows that ChatGPT's scores in linguistics are closer to each other than in mathematics. Hence, ChatGPT is generally more consistent with linguistics. Based on the results of the simulations, ChatGPT is more consistent in linguistics than mathematics, showing its capability to aid high school students in linguistics, especially in tasks related to interpreting and analyzing passages. However, it scores the lowest in the Relevant Words in Context subarea, implicating that ChatGPT cannot successfully discern the context clues or words in sentences.

Meanwhile, in mathematics, it fails to exhibit any clear trend throughout all simulations. This shows that ChatGPT's ability to consistently generate accurate answers in mathematics is still not entirely dependable, at least compared to linguistics.

Lastly, the use of the "Give answer only" prompt makes ChatGPT more prone to giving incorrect responses. This reveals that prompts have an impact on influencing ChatGPT's consistency. This trend is highly exhibited in its responses to mathematics questions in batches 3a and 3b. Table 5 depicts a sample of its responses in each setup for the first question of these two batches. In this question, the correct answer was D. When ChatGPT was prompted to provide its answer only, its choices across all four setups were inconsistent and incorrect. Without the prompt, ChatGPT arrived at the correct answer for all setups. From these results, it can be observed that ChatGPT tends to answer more mathematical questions correctly when it has the liberty to show the thought process it used to arrive at the answer.

Figure 12: Machine B Time A Question No. 29 (with prompt)

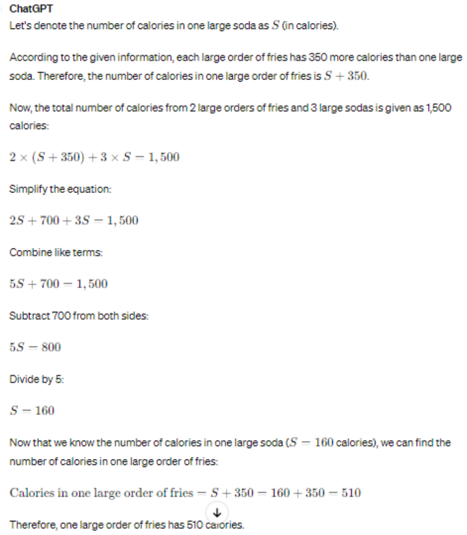


You
At a restaurant, each large order of fries has 350 more calories than one large soda. If 2 large orders of fries and 3 large sodas have a total of 1,500 calories, how many calories does one large order of fries have?

Give the answer only.

ChatGPT
One large order of fries has 650 calories.

Figure 13: Machine B Time A Question No. 29 (no prompt)



ChatGPT
Let's denote the number of calories in one large soda as S (in calories).

According to the given information, each large order of fries has 350 more calories than one large soda. Therefore, the number of calories in one large order of fries is $S + 350$.

Now, the total number of calories from 2 large orders of fries and 3 large sodas is given as 1,500 calories:

$2 \times (S + 350) + 3 \times S = 1,500$

Simplify the equation:

$2S + 700 + 3S = 1,500$

Combine like terms:

$5S + 700 = 1,500$

Subtract 700 from both sides:

$5S = 800$

Divide by 5:

$S = 160$

Now that we know the number of calories in one large soda ($S = 160$ calories), we can find the number of calories in one large order of fries:

$\text{Calories in one large order of fries} = S + 350 = 160 + 350 = 510$

Therefore, one large order of fries has 510 calories.

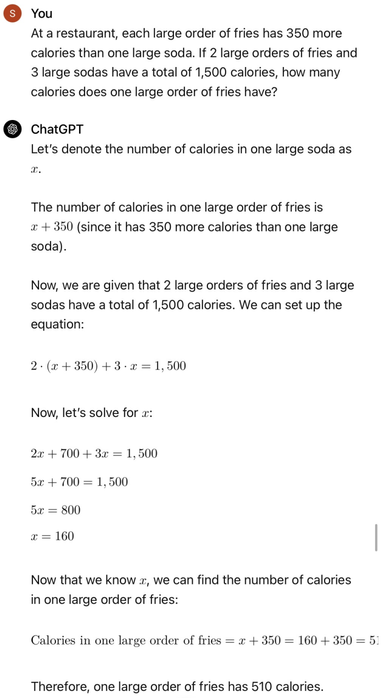
Figures 12 to 14 show sample conversations with ChatGPT. ChatGPT is asked to answer a Mathematics question. In Figure 12, the "Give answer only prompt" is included after posting the question to ChatGPT, answering 650 calories. In Figure 13, the same question is given to ChatGPT, excluding the prompt. ChatGPT answered the problem by showing its solution, obtaining an answer of 510 calories, different from its answer when a prompt was included at the end of the problem.

The conversation shown in Figure 14 is a Temporal and Machine variation setup. For this question, ChatGPT ended up with the same answer and solution as Figure 13, only using different variables. In other cases, ChatGPT may find other ways to approach a problem but still obtain the correct answer.

Tables 6 and 7 provide the number of instances ChatGPT remains consistent with its responses and answered correctly across all setups in both linguistics and mathematics. The following can be observed from the tabulated results:

(1) ChatGPT has a higher frequency of maintaining correct answers than maintaining incorrect answers between the two batches in both linguistics and mathematics.
(2) There is a higher frequency of non-matching answers between the two batches in the domain of mathematics.
(3) In all four setups, ChatGPT gets more mathematics questions correctly in Batch 3b than in Batch 3a. This further suggests that ChatGPT tends to be more accurate when the "Give answer only" prompt is omitted.

Figure 14: Machine C Time C Question No. 29 (no prompt)



S You

At a restaurant, each large order of fries has 350 more calories than one large soda. If 2 large orders of fries and 3 large sodas have a total of 1,500 calories, how many calories does one large order of fries have?

ChatGPT

Let's denote the number of calories in one large soda as x.

The number of calories in one large order of fries is $x + 350$ (since it has 350 more calories than one large soda).

Now, we are given that 2 large orders of fries and 3 large sodas have a total of 1,500 calories. We can set up the equation:

$$2 \cdot (x + 350) + 3 \cdot x = 1,500$$

Now, let's solve for x:

$$2x + 700 + 3x = 1,500$$

$$5x + 700 = 1,500$$

$$5x = 800$$

$$x = 160$$

Now that we know x, we can find the number of calories in one large order of fries:

$$\text{Calories in one large order of fries} = x + 350 = 160 + 350 = 5$$

Therefore, one large order of fries has 510 calories.

Table 6: Comparison of ChatGPT's Answers in Batches 3a and 3b (Linguistics)

	TQ*	MC@	MIC#	≠3a°	≠3b^	≠Inc’
Machine A, Time A	40	16	12	3	5	4
Machine B, Time A	40	19	9	3	4	5
Machine A, Time B	40	17	16	1	2	4
Machine C, Time C	40	15	6	7	5	7

* Total Number of Questions

@ Answers match, both correct

# Answers match, both incorrect

° Answers don't match, correct in Batch 3a

^ Answers don't match, correct in Batch 3b

’ Answers don't match, incorrect in both batches

(4) In instances where ChatGPT is correct in Batch 3a but not in Batch 3b, it is plausible that ChatGPT generated learned the said item. In other words, the question and its answer may have been part of the dataset used by ChatGPT in learning. As for the reason it becomes incorrect in the absence of the prompt, it is likely that ChatGPT follows a structure of problem-solving it has learned and generates a solution using one of these structures.

5 CONCLUSION

Based on the outcome of the simulations, ChatGPT is more consistent in linguistics than in mathematics. It showed its capability in linguistics, specifically in tasks related to interpreting and analyzing passages. However, the results show that ChatGPT cannot consistently discern the context clues or words in sentences. Meanwhile, in mathematics, ChatGPT fails to exhibit any clear trend throughout all simulations. This shows that ChatGPT's ability to consistently generate accurate answers in mathematics is still not entirely dependable. Lastly, utilizing the "Give answer only" prompt makes ChatGPT vulnerable to giving incorrect responses. This reveals that prompts play an important role in influencing ChatGPT's performance.

Table 7: Comparison of ChatGPT's Answers in Batches 3a and 3b (Mathematics)

	TQ*	MC@	MIC#	≠3a°	≠3b^	≠Inc’
Machine A, Time A	40	12	1	2	13	12
Machine B, Time A	40	5	2	3	23	7
Machine A, Time B	40	16	3	2	13	6
Machine C, Time C	40	5	0	2	22	11

* Total Number of Questions

@ Answers match, both correct

# Answers match, both incorrect

° Answers don't match, correct in Batch 3a

^ Answers don't match, correct in Batch 3b

’ Answers don't match, incorrect in both batches

As this study assesses ChatGPT's consistency across different SAT subareas in linguistics and mathematics, future works could determine the performance of ChatGPT on the specific areas e.g. evaluation of simple arithmetic expressions, algebraic expression, geometry-related questions, along with the respective reasons. Moreover, prompts have a significant impact on ChatGPT's performance. Further studies may focus on identifying the proper prompts for ChatGPT to perform with a high level of consistency. A definition of how prompts should be formed for ChatGPT to achieve an ideal performance is also a possible research focus.

REFERENCES

[1] College Board. 2023. SAT program results for the class of 2023 show continued growth in SAT participation. https://newsroom.collegeboard.org/sat-program-results-class-2023-show-continued-growth-sat-participation.

[2] Ian Bogost. 2023. Is this the singularity for standardized tests? https://www.theatlantic.com/technology/archive/2023/03/open-ai-gpt4-standardized-tests-sat-ap-exams/673458/.

[3] Peter Andr'e Busch and Geir Inge Hausvik. 2023. Too good to be true? An empirical study of ChatGPT capabilities for academic writing and implications for academic misconduct. https://www.researchgate.net/publication/370106469_Too_Good_to_Be_True_An_Empirical_Study_of_ChatGPT_Capabilities_for_Academic_Writing_and_Implications_for_Academic_Misconduct. *Twenty-ninth Americas Conference on Information Systems.*

[4] Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. 2023. ChatGPT goes to law school. http://dx.doi.org/10.2139/ssrn.4335905. *Journal of Legal Education* 71 (2023), 387.

[5] Joost C. F. de Winter. 2023. Can ChatGPT Pass High School Exams on English Language Comprehension? https://doi.org/10.1007/s40593-023-00372-z. *International Journal of Artificial Intelligence in Education* (2023).

[6] Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, and Julius Berner. 2023. Mathematical capabilities of ChatGPT. https://arxiv.org/abs/2301.13867.

[7] Robert Lew. 2023. ChatGPT as a COBUILD lexicographer. https://doi.org/10.1057/s41599-023-02119-6. *Humanities and Social Sciences Communications* 10, 1 (2023), 704.

[8] Neville J. McKenzie. March 8, 2023. The limitations and biases of ChatGPT: A critical look. https://www.linkedin.com/pulse/limitations-biases-chatgpt-critical-look-neville-j-mckenzie.

[9] Hannah Muniz. 2021. The 4 SAT sections: What they test and how to do well. https://blog.prepscholar.com/sat-sections.

[10] Erin Ohsie-Frauenhofer. 2023. The SAT will become fully digital—and shorter—by 2024. Here's what's changing and what's staying the same. https://blog.arborbridge.com/sat-will-become-fully-digital-and-shorter-by-2024-whats-changing.

[11] Paulo Shakarian, Abhinav Koyyalamudi, Noel Ngu, and Lakshmivihari Mareedu. 2023. An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP). https://api.semanticscholar.org/CorpusID:257219363. *ArXiv* abs/2302.13814 (2023).

[12] USAFacts Team. 2022. Are fewer students taking the SAT? https://usafacts.org/articles/are-fewer-students-taking-the-sat/.

[13] Christian Terwiesch. 2023. *Would Chat GPT3 get a Wharton MBA? A prediction based on its performance in the operations management course.* Technical Report. William and Phyllis Mack Institute for Innovation Management. https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP.pdf.

[14] UNESCO. 2024. Sustainable development goal 4 (SDG 4).

[15] Yousef Wardat, Mohammad A. Tashtoush, Rommel AlAli, and Adeeb M. Jarrah. 2023. ChatGPT: A revolutionary tool for teaching and learning mathematics. https://doi.org/10.29333/ejmste/13272. *Eurasia Journal of Mathematics, Science and Technology Education* 19, 7 (2023).