Manila Journal of Science 17 (1) 2025, pp. 12-36

GPT Gets Quizzed: A Mixed-Method Study on ChatGPT's Performance in Answering the SATs

Eliza Janel L. Tan[1]*, Keith Alexandre L. Ramos[1], Maria Elisha Kaye B. Nazario[1], Steven Vincent D. Lim[1], and Shirley B. Chu[2]
[1]Science, Technology, Engineering, and Mathematics Department, De La Salle University Integrated School, Manila, Philippines
[2]College of Computer Studies, De La Salle University, Manila, Philippines

*Corresponding Author: eliza_janel_tan@dlsu.edu.ph

ABSTRACT

Artificial intelligence has been integral to daily societal systems, including modern education through tools like OpenAI's ChatGPT. While past studies have assessed ChatGPT's performance in various domains, such as law and medicine, a gap in research on the analysis of its efficacy in secondary school-level subjects persists. Therefore, this study assessed the performance of ChatGPT in high school level linguistics and mathematics questions, in correlation to the perceptions of students and professors. By extension, it provides a more detailed analysis of ChatGPT's potential as a learning tool. To achieve this, SAT questions are administered to ChatGPT. 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. Additionally, ChatGPT was also observed to perform better than at least 50% of high school SAT student test takers, with accuracy rates of 59.59% in linguistics and 56.41% in mathematics. Through survey and interview, the study also reveals that there is a gap between student perception on ChatGPT's performance than its simulation accuracy rate. In linguistics, there was a significant gap between the mean survey with interview results and the simulation accuracy, while in mathematics, the gap was smaller.

Keywords: ChatGPT, SATs, accuracy, consistency, learning tool, high-school, linguistics, mathematics

INTRODUCTION

ChatGPT was introduced by the company OpenAI to the world in November 2022. It was made to respond to inquiries or prompts provided by human users using information it gathers and stores from datasets up to September 2021. If it is asked to, ChatGPT can modify its responses using human feedback.

Since the day that it was released to the public, it has become one of many artificial intelligence (AI) softwares that has benefitted the education sector, helping both learners and educators. Students have used ChatGPT for assistance in schoolwork across different academics. Baker McKenzie (2023) also notes that ChatGPT has built its reputation among high school students as a learning tool as it aids them in comprehending topics under the subject domains of mathematics and linguistics. It does so by providing increased access to information and personalized learning as per Li and Xing (2021), Farrokhnia et. al (2023), and Fuchs (2023). Despite the benefits that ChatGPT has provided students in guiding them through their learning journeys, there are potential risks that students might become increasingly dependent on it and that it could unintentionally provide false information. (Rogerson, 2023). The integration of ChatGPT in education may elicit mixed perceptions as highlighted by Bonsu and Baffour-Koduah (2023) and Mohamed (2023). However, a study by Firat (2023) highlights key themes and sentiments observed by interviewed students and professors, notably evolution of open and distance learning, rethinking assessment methods, and changing role of educators, and social and ethical concerns. These themes indicate that students and teachers alike believe that ChatGPT will have a significant impact on traditional learning paradigms

Although there were already studies that attempted to quantitatively analyze ChatGPT's proficiency in various subject domains, such as Terwiesch (2023) and Choi et al. (2023) investigated its performance in answering assessments, it was identified that there is still the need to investigate how ChatGPT would perform when it is tasked to answer high-school level questions, especially those under mathematics and linguistics, to determine if it is an adequate tool that high school students can use to comprehend these subjects. In mathematics, past studies like Frieder et al. (2023) has shown that while ChatGPT proves its knowledgeability on commonly known concepts and theorems, it was concluded that its answers may be partially correct or completely incorrect due to system errors in algebra, computation, and logical flow of presented arguments. In linguistics, De Winter (2023) using the national examinations of the Netherlands's VWO (Preparatory Scientific Education) program for English reading comprehension, revealed that ChatGPT performed as proficiently as the average student, even if GPT-4 outperformed both of them.

Thus, the researchers chose to study ChatGPT's performance in answering mathematics and linguistics questions from the College Board's SATs, formerly known as the Scholastic Aptitude Test, which is a standardized examination taken by over a million students worldwide annually to gauge their readiness for college (College Board, 2023; Muniz, 2021; USAFacts Team, 2022). Considering that the College Board plans to implement the admission of digital SATs, it is highly likely that students will use ChatGPT to prepare for these standardized examinations to get higher scores (Bogost, 2023; Ohsie-Frauenhofer, 2023). Therefore, there is the need to assess ChatGPT's accuracy and consistency in answering high-school level questions to determine its effectiveness as a tool to aid students in preparing for tests. Investigating and understanding the interconnection between ChatGPT and

the SATs also presents the opportunity for researchers to examine ChatGPT's performance in answering questions and compare it to human test takers. At the same time, academic stakeholders including students, professors, and parents would be able to make more informed decisions with regard to using ChatGPT for school.

This research aims to assess the accuracy and consistency of ChatGPT's answers to high-school level mathematics and linguistics questions and determine ChatGPT's effectiveness as a learning tool through the perspectives of academic stakeholders. To achieve this, the study uses a mixed-method approach to attain qualitative and quantitative data on ChatGPT's capabilities and limitations within these subject domains. The researchers will administer sample SAT linguistics (gathered from both the Reading and the Writing and Language SAT sections) and mathematics questions to ChatGPT, collected from online and offline practice tests, and evaluate its performance in comparison to that of a high school student. Along with this, the researchers will also collect the perspectives of academic stakeholders from De La Salle University - Manila Campus on the strengths and possible performances of ChatGPT in this standardized examination. This will be done by administering an online survey to senior high school students in the STEM, ABM, and HUMSS strands and conducting face-to-face interviews with professors handling senior high school linguistics and mathematics subjects. learning tool. Although similar methods have been utilized in previous studies, this study provides a more updated review of student and professor perception towards the implementation of ChatGPT as a tool in the education system. At the same time, this study attempts to see whether or not there is a gap in the perceived accuracy of ChatGPT and its actual accuracy in linguistics and mathematics.

The research specifically aims to accomplish the following:

1. To assess and compare the accuracy rate and consistency of ChatGPT in the linguistics (Reading and Writing & Language) and mathematics sections of SAT practice tests gathered both online and offline;
2. To analyze the performance of ChatGPT in answering objective questions in linguistics and mathematics;
3. To compare the SAT score of ChatGPT in linguistics and mathematics to the performance of different genders, race, and overall student test takers;
4. To evaluate the perceptions of selected DLSU-Manila students and educators on ChatGPT's objective accuracy in linguistics and mathematics in relation to student learning.

Scope and Limitations

ChatGPT's performance is measured by comparing its answers to selected datasets of linguistics and mathematics SAT practice test questions gathered from online and offline sources. Only objective questions, that is, questions with a fixed answer are included in the dataset. These objective questions include multiple-choice linguistics questions, multiple-choice mathematics questions, and free-response mathematics questions.

This study only covers understanding the performance of the GPT-3.5 model of ChatGPT. At the time when the surveys were conducted, the GPT-3.5 model was accessible for free to high school students and educators, while the multimodal version GPT-4 was not.

Since ChatGPT cannot process and respond to prompts that include images and geometric figures, questions from the mathematics practice tests involving these are not included.

MATERIALS AND METHODS

Research Design

The study adopts a mixed-method triangulation approach. The quantitative aspect involves the ChatGPT simulation that involves feeding ChatGPT SAT questions and comparing its answers to the expected ones, while the qualitative aspect describes the execution of surveys and interviews to supplement and relate the results of the simulation to user perceptions among students and professors. The entire procedure is illustrated in Figure 1.

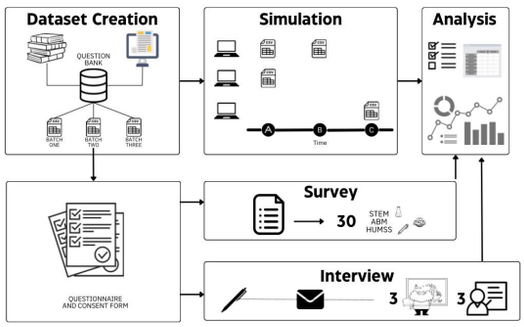

Figure 1: Schematic Representation of Study's Methodology

Dataset Creation

To prepare for the simulations with ChatGPT, the dataset collection phase involves the creation of a question bank to categorize and store SAT questions for linguistics and mathematics sets from different online and offline sources and record ChatGPT's response to each question. The questions are categorized by source to determine if ChatGPT performs better in answering questions gathered from practice SATs on the Internet or questions gathered from a printed book.

The researchers hypothesize that during the simulation, it would perform better in answering questions gathered from online sources. This is because ChatGPT is trained to search for patterns in words used in prompts and spit out strings of text, from the data used to train it, that it assumes to be the correct answer for these questions (Guinness, 2024). This set of data that ChatGPT uses to answer questions comes from sources found all around the Internet. Therefore, it is assumed that if ChatGPT sees a question that has already been on an online source, then it would answer this question accurately and consistently.

Questions are categorized and labeled by source and subject. These datasets are randomized and arranged into three batches of 40 questions per domain. Each set is divided equally among the domain's subareas. For 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**: emphasizes the structures of sentences, usage, and punctuation;
2. **Expression of Ideas:** emphasizes topic development, construction, and rhetorically effective use of language;
3. **Relevant Words in Context:** emphasizes inscribing word/phrase meaning in context and rhetorical word choice;
4. **Command of Evidence:** emphasizes how to use and understand the material presented in sections and informational graphics (e.g. tables, charts, and graphs).

For mathematics, questions encompass multiple types: multiple choice, free-response, equation, and tables. Similarly, ten questions from each of the following subareas in the mathematics section form each batch of dataset:

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

Questions were categorized and labeled according to source and subject, ensuring an equal representation of offline and online sources for a fair comparison. Since input ChatGPT 3.5 is restricted to text only, some questions are modified. In linguistics, additional instructions are included for revising the underlined words or phrases in the original texts. For mathematics, LaTeX was used for typesetting mathematical expressions, enabling input of complex symbols and structures. Examples of revisions are shown in Table 1.

Simulation

The simulation proper involves querying ChatGPT with questions from the dataset and recording its responses. Inspired by methodologies in studies by Shakarian et al. (2023), Wardat et al. (2023), and De Winter (2023), ChatGPT's consistency in a subarea is examined by having it answer each question four times under different conditions or on different machines. The setups 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 2 for a comparison of independent variables per setup.

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

LINGUISTICS		MATHEMATICS (Equation)		MATHEMATICS (Table)	
Original	**Input to ChatGPT**	**Original**	**Input to ChatGPT**	**Original**	**Input to ChatGPT**
Nutritionists consider Greek yogurt to be healthy food: it is an excellent source of calcium and protein, serves to be a digestive aid, and it contains few calories in its unsweetened low- and non-fatfat forms. A. NO CHANGE B. containing C. contains D. will contain	Nutritionists consider Greek yogurt to be healthy food: it is an excellent source of calcium and protein, serves to be a digestive aid, and it contains few calories in its unsweetened low- and non-fatfat forms. **What improvements can be made to"it contains"?** A. NO CHANGE B. containing C. contains D. will contain	The length of a rectangle's diagonal is $5\sqrt{17}$, and the length of the rectangle's shorter side is 5. What is the length of the rectangle's longer side? A) $\sqrt{17}$ B) 20 C) $15\sqrt{2}$ D) 400	The length of a rectangle's diagonal is **5\sqrt{17}**, and the length of the rectangle's shorter side is 5. What is the length of the rectangle's longer side? A) **\sqrt{17}** B) 20 C) **15\sqrt{2}** D) 400	The bar graph above shows the number of students in four high schools who play on a school sports team. The number of students enrolled in schools A, B, C, and Dare 1,500, 2,800, 550, and 3,300, respectively. Scho ol / Num ber of Stud ents Who Play on a Sport s Team A 600 B 1000 C 200 D 1200 In which school do the highest percent of the students play on a team? (A) A (B) B (C) C (D) D	**\begin{table} [] \begin{tabular}{llllll} School & Number of Students Who Play on a Sports Team & & \\ A & 600 & & \\ B & 1000 & & \\ C & 200 & & \\ D & 1200 & & \\ & & & \end{tabular} \end{table}** The bar graph above shows the number of students in four high schools who play on a school sports team. The number of students enrolled in schools A, B, C, and Dare 1,500, 2,800, 550, and 3,300, respectively. In which school do the highest percent of the students play on a team? (A) A (B) B (C) C (D) D

Table 2: Independent Variables per Setup for Comparison

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

During the simulation, a conversation with ChatGPT is initiated. ChatGPT is asked to answer 40 SAT questions for each domain, one at a time, during this conversation. ChatGPTs

answer to each question is then recorded. In some instances, additional prompts are added after each question. In Batch One, the prompt "Give answer only" was added to some of the questions in both linguistics and mathematics. In Batch Two, no such prompt was given. In Batch Three, the prompt was added to all questions. Conversation samples are shown in Figures 2 to 4.

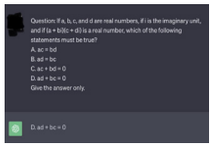

Figure 2: Conversation Sample of ChatGPT to a Multiple Choice Mathematics Question

Figure 3: Conversation Sample of ChatGPT to a Linguistics Question

Figure 4: Conversation Sample of ChatGPT to a Free Response Mathematics Question

There are instances when ChatGPT's answer to the same question is inconsistent, that is, different letter or numerical responses given with the answer only prompt and without. For instance, in Figure 5, ChatGPT was tasked to answer a multiple-choice mathematics question where the answer was choice C (22). When it was given the question with the "give the answer only" prompt, it incorrectly selected choice B (1). However, when it was asked immediately afterward to show its process for solving the problem, it showed the complete and correct solution while at the same time changing its response to choice C. In cases like these, its initial answer is documented.

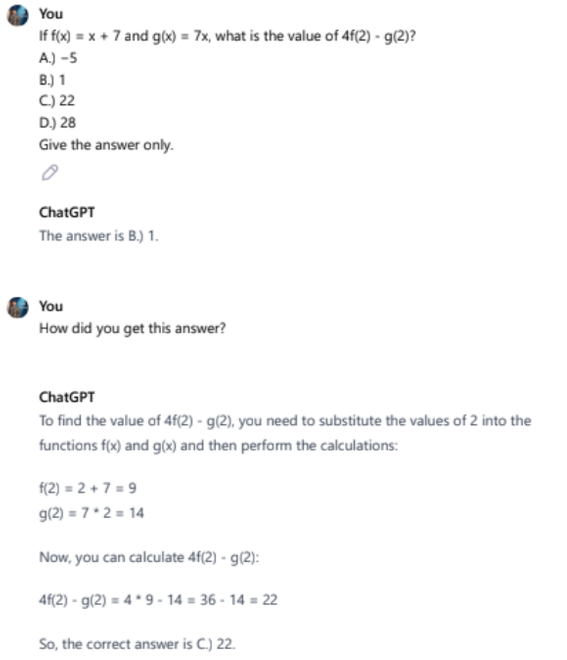

Figure 5: Sample of ChatGPT's Inconsistent Responses to a Single Question

Data Collection

The quantitative data are derived from the results of the simulations. For every question in which ChatGPT's answers match the expected answer, a point is awarded and added to its raw score .

Table 3. Variables and Symbols to be used in the Formulae

Variable	Symbol
Objective Accuracy Rate in Subarea	$A_{subarea}$
Objective Accuracy Rate in Domain	A_{domain}
Estimated Score of ChatGPT in a Domain in a Simulation	ES_{domain}
Maximum Score Possible in a Particular Domain Based on the SAT Conversion Table	$\max(S_{SAT\ Table})$
Raw Score of ChatGPT in a Simulation	S_{raw}
Mean Raw Score of ChatGPT in a Simulation	$\overline{S_{raw}}$
Maximum Score Possible in a Particular Domain Based on the SAT Conversion Table	$\max(S_{domain})$

The objective accuracy is interpreted as ChatGPT's ability to provide the expected answer. This is computed for each subarea of

the linguistics and mathematics domain (Eq. 1), and for each of the domains (Eq. 2).

$$A_{\text{subarea}} = \left(\frac{S_{\text{raw}}}{10}\right) \times 100 \quad \text{Eq. 1}$$

$$A_{\text{domain}} = \left(\frac{S_{\text{raw}}}{40}\right) \times 100 \quad \text{Eq. 2}$$

To determine the performance of ChatGPT in each domain as compared with human test takers, the raw scores for each domain is first converted using the appropriate conversion factors (Eq. 3):

$$ES_{\text{domain}} = \max(S_{\text{SAT Table}}) \frac{\overline{S_{\text{raw}}}}{\max(S_{\text{domain}})} \quad \text{Eq. 3}$$

The raw score $\overline{S_{raw}}$ is the mean score of all setups in the specific domain. Linguistics is divided into reading and writing to adhere to the provided SAT score computation. After the accuracy rate , say is computed, the SAT score calculator is used to convert this raw score to the SAT-scaled score for reading. The same is true for writing and mathematics.

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.

In the survey, participants were asked to predict the score of ChatGPT if provided with 10 questions for each subarea. The mean of these predicted scores for each subarea is computed. In order to have an idea of the respondents' perceptions about ChatGPT, the mean scores are compared with the corresponding actual raw score of ChatGPT.

RESULTS AND DISCUSSION

This chapter presents four major parts. A portion of the data are from the researchers' past papers namely "AI to The Test: Measuring ChatGPT's Objective Accuracy in Answering the SATs in Comparison to Human Performance" by Tan et al. (2024) and "AI Skill Check: An Examination of ChatGPT's Consistency in Linguistics and Mathematics using the SAT" by Nazario et al. (2024).

The first part provides analysis of ChatGPT's accuracy in linguistics and mathematics respectively. This includes the identification of its strengths and weaknesses among the SAT subareas as well as a review of the influence of the prompts "Give answer only" on its performance. The second part explores ChatGPT's consistency in answering the prompted linguistics and mathematics questions of the SATs. This includes understanding the domain when it is most and least consistently correct. The last two sections relate the gathered accuracy and consistency to the predictions of the student and professor population of De La Salle University Integrated School - Manila and De La Salle University to better understand whether or not there is a gap between the perceived accuracy rates and strengths to the actual results. Their perceptions are analyzed through thematic analysis, which involves identifying common ideas and themes regarding ChatGPT's performance in their responses.

While this chapter focuses on comparing the performance of ChatGPT across each of the four setups, it will not tackle the effects of changing the variables such as location and time of the simulation. This is because the simulation showed no conclusive results that changing the independent variables affected ChatGPT's overall performance after each batch.

4.1 Accuracy Results

In both domains, time and machine seem to insignificantly affect the performance of ChatGPT. Hence, it is not thoroughly discussed.

Overall Accuracy of ChatGPT in Linguistics

It was observed that with the "Give answer only" prompt in Batch Three (a), ChatGPT ended up getting more questions incorrectly as compared to the other batches. A second simulation, Batch Three (b) was performed with this additional prompt omitted. To ensure consistency, Batch Three (a)-'s results are excluded in computing the overall mean accuracy rate of ChatGPT. Nevertheless, ChatGPT's overall accuracy rate stands at **59.59%.** Table 4 shows ChatGPT's achieved accuracy for all batches.

Table 4. Mean Accuracy of ChatGPT in Linguistics in Each Batch

	Batch One	Batch Two	Batch Three(a)	Batch Three (b)
Overall	56.23%	73.13%	50.63%	49.38%

Overall Accuracy of ChatGPT in Mathematics

The overall batch accuracy of ChatGPT in the SAT Mathematics stands at **56.41%,** with Batch Three(a) results excluded. This is 3.18% below its performance in Linguistics. Table 5 provides the computed mean accuracy rates.

Table 5. Mean Accuracy of ChatGPT in Mathematics in Each Batch

	Batch One	Batch Two	Batch Three(a)	Batch Three(b)
Overall	55.63%	72.50%	31.25%	66.25%

Mathematics vs. Linguistics

In Batch One and Batch Two, ChatGPT performed more accurately in linguistics than in mathematics. The higher accuracy rate in linguistics may stem from ChatGPT's primary design as a language chatbot rather than a specialized Mathematics tool.

However, this trend only seems to apply to Batch Three(a) (50.63%). In Batch Three(b), ChatGPT achieved a higher accuracy rate in Mathematics (66.25%) over Linguistics (49.38%). This is likely due to the influence of the omission of the prompt and the re-feeding of the questions to ChatGPT.

The Effects of "Give answer only" Prompt

In Batch Three (a), ChatGPT scored a lower objective accuracy rate of 31.25% compared to the first two batches (55.63% and 72.50% for Batch One and Batch Two, respectively). However, in Batch Three (b), its overall objective accuracy rate increased to 66.25%. This may demonstrate that ChatGPT attains higher objective accuracy in Mathematics when it is not limited to just giving an answer. Similarly to how a human test taker would answer a mathematics question, ChatGPT also is more likely to obtain the correct answer when it shows its thought process or solution compared to giving an answer only.

Linguistics Subarea Performance

Table 6 shows ChatGPT's mean accuracy for each subarea under linguistics.

Table 6. Mean Accuracy of ChatGPT

	Command of Evidence	Relevant Words in Context	Expression of Ideas	Standard English Conventions
Mean	80.00%	49.38%	55.00%	45.00%

ChatGPT has the highest accuracy in Command of Evidence. It has low accuracies in Relevant Words in Context and Standard English Conventions. This is consistent with De Winter's (2023) observation that relative to other linguistics-related tasks, ChatGPT is most capable in noting details

from texts to identify information through reading comprehension skills (Command of Evidence). ChatGPT has the lowest accuracy in addressing word/phrase meaning in context and rhetorical word choice (Relevant Words in Context); and sentence structure, usage, and punctuation (Standard English Conventions).

Mathematics Subarea Performance

Table 7 shows ChatGPT's mean accuracy for each subarea under mathematics.

Table 7. Mean Accuracy of ChatGPT in Each Mathematics Subarea

	Heart of Algebra	Problem Solving & Data Analysis	Passport to Advanced Math	Additional Topics in Math
Mean	56.88%	50.63%	60.63%	57.50%

From the mean accuracy rates in the Mathematics domain (Table 5), ChatGPT's accuracy is highest in Passport to Advanced Math (60.63%) and lowest in Data Analysis and Problem Solving (50.63%). This indicates that ChatGPT is least inaccurate at interpreting data (Data Analysis), while its accuracy is high in analyzing evidence from text (Command of Evidence). Additionally, it is capable of reasoning with more complex equations, and interpreting and building functions (Passport to Advanced Math).

Comparison to Human Performance

Now, the performance of ChatGPT is compared to the actual performance of high school students who took the SATs in 2022. Table 8 provides the ChatGPT's mean raw scores of each batch as well as the mean overall score, which was used as a basis to get ChatGPT's SAT scores once scaled.

Table 8. Mean Scores of ChatGPT

	Mean # of Correctly Answered Questions		
SETUP	Reading (/20)	Writing (/20)	Mathematics (/40)
Batch One	12.00	10.50	22.25
Batch Two	17.50	12.00	29.00
Batch Three (b)	11.50	8.25	26.50
Mean	13.67	10.25	25.92

In order to compare ChatGPT's performance on the SAT with human test takers, scores that closely resemble actual SAT results were needed. Keyman (2018) found that Practice Test 6 scores best reflect real SAT performance. Therefore, the raw scores obtained were converted to SAT scores using the conversion factors used in Practice Test 6. To compare ChatGPTs performance with human test takers, ChatGPT's score is compared to that of human test takers, using the data from the 2022 SAT Suite of Assessments Annual Report.

Using Equation 3, ChatGPT's calculated scores in each SAT section are approximately 35 points out of 52 for Reading, 22 points out of 44 for Writing, and 37 out of 58 for Mathematics. Upon inputting these individual scores into the SAT Score Calculator, it is determined that ChatGPT garnered a total score of 1130 points.

Notably, ChatGPT has a higher score in Mathematics than in Linguistics, even if it has a higher mean raw score in Linguistics. This may be due to the way Linguistics scores are computed. It is apportioned into the Reading, and Writing & Language components in the SATs.

Based on the collated SAT mean scores of high school seniors who took the SAT, ChatGPT generally scored higher than human students, scoring 1130 to the student mean score of 1050.

In linguistics and mathematics, ChatGPT performed better than at least half of the males and the females who took the SATs. However, relative to different races and ethnicities, ChatGPT scored lower compared to at least 50% of test takers who identify themselves

as White, Asian, and Mixed respectively in linguistics. In mathematics, it scored lower than 50% of Asians who took the SATs.

Nevertheless, compared to an average high school student in an aggregate perspective, ChatGPT exhibits more proficient linguistics- and mathematics-related abilities. Hence, to a certain extent, ChatGPT may be effective as a learning tool for at least 50% of high school students in understanding concepts relevant to high school linguistics and mathematics.

4.2 Consistency Results

Consistency is defined as ChatGPT's tendency to generate the same response when given the same prompt. Hence, the term *consistently correctly* refers to instances where ChatGPT answers the same correct response across all setups. Conversely, *consistently incorrectly* describes instances where ChatGPT repeatedly generates the same incorrect response (e.g. it answers B across all setups when the correct answer is A). Meanwhile, the term inconsistent will be used to describe instances where ChatGPT generates different responses in the four setups.

Overall Consistency of ChatGPT in Linguistics

The summary of ChatGPT's consistency in each batch in linguistics is shown in Table 9. In Batch 1 and Batch 2, ChatGPT follows a consistently correct trend. There is even a minor improvement in ChatGPT's rate in getting consistently correct answers in Batch 2. This consistently correct trend could be a result of ChatGPT being able to recognize the text patterns in linguistic questions and match these patterns to its training data as it generates an answer. Moreover, it is possible that SAT linguistics questions from online sources are plentiful in ChatGPT's training data, leading to ChatGPT being more familiar with the correct answers and explanations to questions.

However, by Batch 3a and 3b, ChatGPT starts to become more inconsistent. This inconsistency could be attributed to the set of questions used as 3a and 3b share the same set of questions. It is possible that Batch 3's set of questions have a higher difficulty than the previous batches or that some questions are not part of ChatGPT's training data. Moreover, it is apparent that the "Give answer only" prompt causes no significant change in ChatGPT's consistency in linguistics. This is the case as linguistics questions focus less on step-by-step reasoning and more on mastery of grammatical rules, reading comprehension, and language patterns. Therefore, the presence of a "Give answer only" prompt doesn't remarkably interfere with the rules or understanding that ChatGPT uses in generating its answers.

Table 9. Consistency of ChatGPT's Answers Across Setup in Each Batch (Linguistics)

SETUP	Total Number of Questions	Answers match, all correct (consistently correct)	Answers match, all incorrect (consistently incorrect)	Answers don't match (inconsistent)
Batch 1	**40**	**15**	**11**	**14**
Batch 2	**40**	**20**	**6**	**14**
Batch 3a (with "Give answer only")	**40**	**13**	**6**	**21**
Batch 3b (without "Give answer only")	**40**	**14**	**4**	**22**

Consistency for each Linguistics Subarea

The following section tallies how consistent ChatGPT performs in each of the linguistics subareas for each batch.

Figure 6 shows that in Batch 1, ChatGPT is most consistently correct in Command of Evidence. 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.

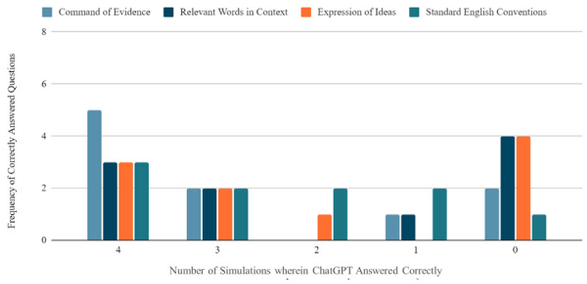

Figure 6. Batch 1 Results (Consistency in Linguistics)

Figure 7 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 more questions consistently correctly compared to 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 results of Batch 1 could be due to a different set of questions being used or to a possible refinement in ChatGPT's ability in conveying thoughts and identifying grammatical errors. Lastly, in all simulations, ChatGPT answers with the fewest questions correct in the Relevant Words in Context subarea.

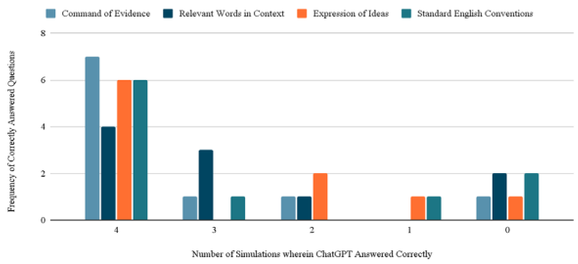

Figure 7. Batch 2 Results (Consistency in Linguistics)

Figure 8 shows Batch 3a results. The "Give answer only" prompt is included in all questions in this simulation. With the addition of the prompt, ChatGPT still maintains the highest consistency of correctly answered questions in the subarea of Command of Evidence, however, 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

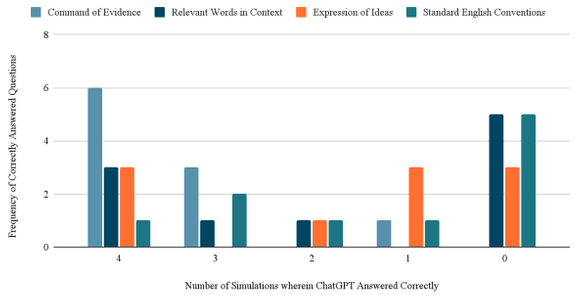

Figure 8. Batch 3a Results (Consistency in Linguistics)

When testing the same set of questions without the "Give answer only" prompt, there is no significant difference in ChatGPT's consistency with regards to best and worst performing subareas (Figure 9). 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. It does not necessarily mean that ChatGPT performs better without it, as that is beyond the scope of this paper.

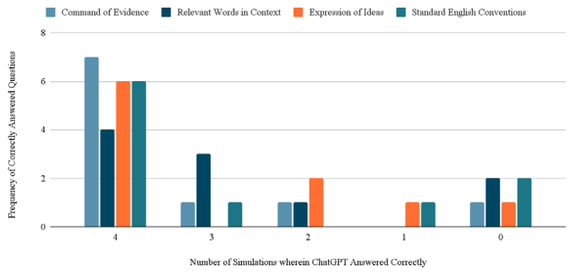

Figure 9. Batch 3b Results (Consistency in Linguistics)

Overall, for linguistics it can be inferred that 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 Wardat et al. (2023). Conversely, 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 differs from observations of Lew (2023) which 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.

Overall Consistency of ChatGPT in Mathematics

ChatGPT's consistency in mathematics is illustrated in Table 10. The inconsistent trend exhibited by all batches could be due to limitations of autoregressive large language models (LLMs) and the heavy step-by-step reasoning of mathematics questions. To reiterate, autoregressive LLMs such as ChatGPT, generate a response starting with an input and predict the next words based on patterns from its training data (Crabtree, 2023). According to Cobbe et al. (2021), ChatGPT's autoregression makes it struggle in multistep reasoning as once it makes an error in one line of solution, it is difficult for it to correct itself. Furthermore, large-language models specialize in generating human-like text based on patterns rather than doing math calculations (Muehmel, 2023).ChatGPT's "next word prediction" mechanism may not be efficient in solving math problems wherein answers are heavily dependent on given values, different wordings, and multi-step solutions (Yavuz, 2024; Zvornicanin, 2024).

Nevertheless, with the exception of Batch 3a, succeeding batches show some improvement in answering consistently correctly. Batch 2 exhibits greater consistency in answering more questions correctly as compared to Batch 1. Meanwhile, Batch 3b also answered more questions consistently correct than Batch 1 but less than Batch 2. This occurrence could be a result of ChatGPT's training data updating in between batches or the varying difficulties of the SAT questions used in each batch.

ChatGPT is the least consistently correct and most inconsistent when the "Give answer only" prompt is added to the questions. This

substantial decline could be a result of the prompt limiting ChatGPT's ability to generate a reasoning or solution, hence, limiting it to simply retrieving a single answer from its database or generating one based on limited context.

Table 10. Consistency of ChatGPT's Answers Across Setup in Each Batch (Mathematics)

(Mathematics)

SETUP	Total Number of Questions	Answers match, all correct (consistently correct)	Answers match, all incorrect (consistently incorrect)	Answers don't match (inconsistent)
Batch 1	**40**	**6**	**1**	**33**
Batch 2	**40**	**18**	**3**	**19**
Batch 3a (with "Give answer only")	**40**	**2**	**2**	**36**
Batch 3b (without "Give answer only")	**40**	**15**	**2**	**23**

Consistency for each Mathematics Subarea

The following section shows ChatGPT consistency in each of the mathematics subareas for each batch.

It can be seen in Figure 10 that in Batch 1, ChatGPT does not have a pattern in terms of consistency in any of the mathematics subareas. ChatGPT performs the best in Data Analysis & Problem-Solving albeit only getting three out of ten questions consistently correct. 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.

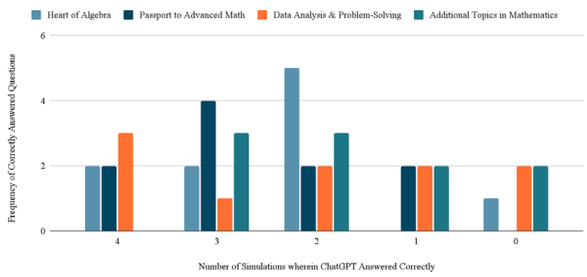

Figure 10. Batch 1 Results (Consistency in Mathematics)

Figure 11 shows that ChatGPT exhibits significant improvement in all subareas in Batch 2 in terms of being consistently correct. Data Analysis and Problem Solving continues to be its best performing subarea. It has the largest improvement in Additional Topics in Mathematics. This could mean that ChatGPT may have been updated in between the Batch 1 and 2 simulations to answer such questions better or that the questions in Batch 2 under Additional Topics are easier or more familiar to ChatGPT.

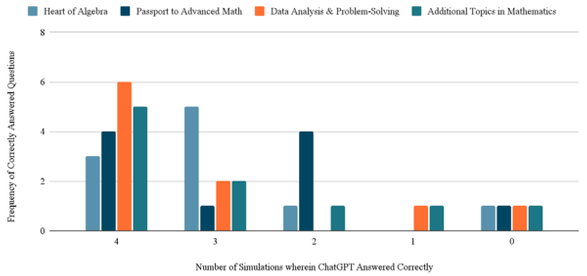

Figure 11. Batch 2 Results (Consistency in Mathematics)

As mentioned previously, the results of Batch 3a (Figure 12), 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*.

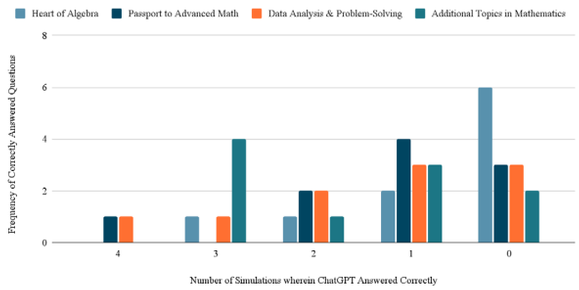

Figure 12. Batch 3a Results (Consistency in Mathematics)

Without the "Give answer only" prompt, ChatGPT's consistency returns to a trend similar to Batches 1 and 2. Figure 13 shows that ChatGPT is now the most *consistently correct* in the Passport to Advanced Math subarea. Notably, results in this simulation show a decline in performance compared to Batch 2.

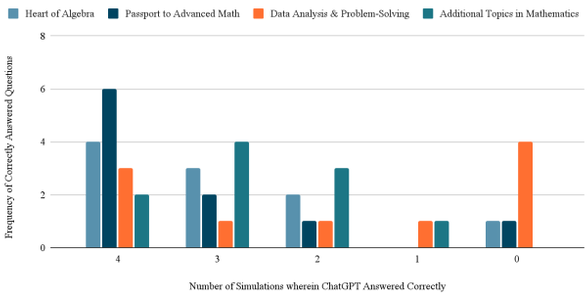

Figure 13. Batch 3b Results (Consistency in Mathematics)

After plotting how often ChatGPT answers consistently correctly or consistently incorrectly for the mathematics simulations, it is observed that ChatGPT does not exhibit any clear trend regarding consistency and inconsistency in the mathematics subareas. For instance, Data Analysis and Problem Solving could be the most *consistently correct* subarea for Batch 1, but be most consistently incorrect by Batch 3b. This supports the findings of de Winter (2023), which posits that ChatGPT struggles to do accurate calculations consistently and falls short compared to chatbots specifically trained to do mathematics.

The Effect of the "Give answer only" Prompt on ChatGPT's Consistency

As observed in the simulations, while the "Give answer only" prompt does not affect ChatGPT's consistency in linguistics, it makes the chatbot more prone to being inconsistent in mathematics. This trend is highly exhibited in its responses to mathematics questions in Batches 3a and 3b. Table 11 presents the responses of each setup in Batch 3a and Batch 3b for Mathematics Question 1. Figures 13 to 17 shows a sample conversation of ChatGPT answering the same question.

Table 11: ChatGPT's Answers to Mathematics Question 29 in Batches 3a and 3b

Question Number 1 **Expected Answer: D**				
ChatGPT's Answer	Control Setup	Temporal Variation	Machine Variation	Machine and Temporal Variation
Batch 3a (Give answer only)	**C**	**B**	**A**	**B**
Batch 3b (liberty to show solution)	**D**	**D**	**D**	**D**

Comparing ChatGPT's responses across all setups in Batches 3a and 3b (Table 5), it is observed that it provides a different incorrect response with the "Give answer only" prompt. Once it was tasked to give both its solution and response, all setups took the same approach to solving the problem and gave the correct answer. These results further illustrate how ChatGPT tends to answer more mathematical questions more consistently correctly when it has the liberty to show the thought process it used to arrive at the answer. It is plausible that in Batch 3b, giving it no limitation likely allowed it to do a more thorough search through its knowledge base, leading it to consider alternatives before ultimately arriving at the expected answer. Nevertheless, it must be noted that its overall performance

may change if a different prompt is used to ask ChatGPT not to show any solution.

In Figure 14 (Temporal Variation) and Figure 15 (Location and Time Variation), the "Give answer only" prompt is included after posting the question to ChatGPT, answering A and B respectively - both of which are incorrect. In Figure 16, the same question is given to a Temporal Variation setup, excluding the prompt. After showing its solution, ChatGPT obtained the answer D, which was the correct answer. In the Location and Time Variation setup without the prompt (Figure 17), ChatGPT still ends up with the same answer and even the same solution as Figure 16. It can be surmised that without the "Give answer only" prompt, ChatGPT can better recognize the math problem as something it has encountered in its training data. Such recognition therefore allows it to apply the information it has learned from its dataset to its response.

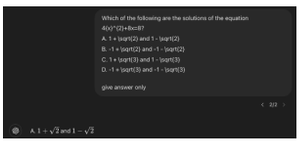

Figure 14. Machine A Time B Question No. 1 (with prompt)

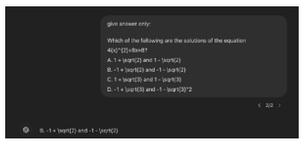

Figure 15. Machine C Time C Question No. 1 (with prompt)

Figure 16. Machine A Time B Question No. 1 (no prompt)

Figure 17. Machine C Time C Question No. 1 (no prompt)

In other cases, ChatGPT may use different variables or have a different flow in answering a problem yet still obtain the correct answer as demonstrated in Mathematics Question No. 29 (Figure 18 and Figure 19). This could be due to its inherent randomness, causing it to generate other possible variables or solutions, especially the less frequent ones, from its database (Wolfram, 2023; Barros, 2023).

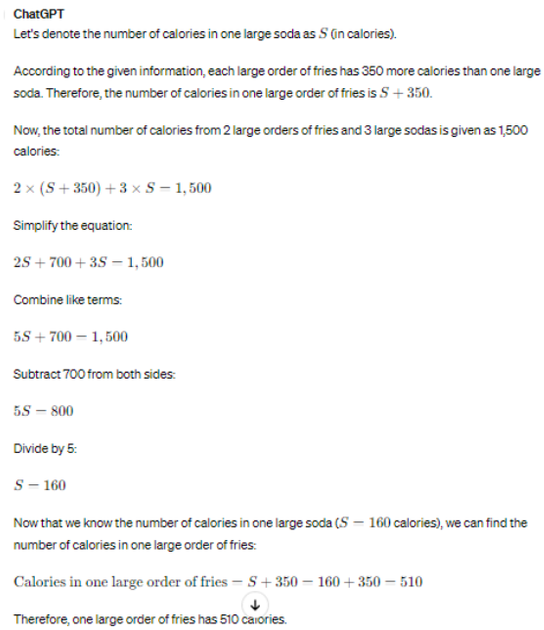

Figure 18. Machine B Time A Question No. 29 (no prompt)

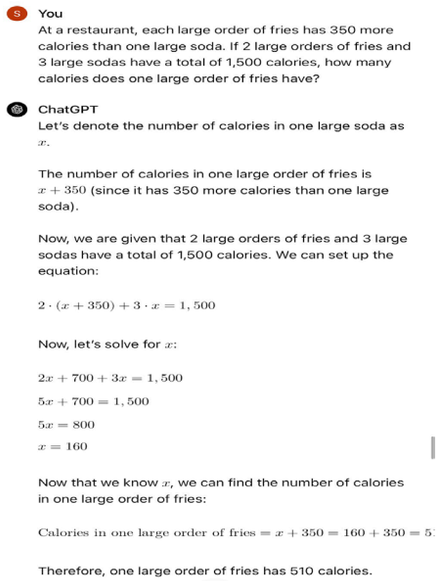

Figure 19. Machine C Time C Question No. 29 (no prompt)

To further elaborate and quantify the effect of the "Give answer only" prompt on ChatGPT's responses in Batches 3a and 3b, Tables 12 and 13 provide the number of instances that it stayed consistent with its responses and answered correctly across all setups in both linguistics and mathematics.

Table 12. Comparison of ChatGPT's Answers in Batches 3a and 3b (Linguistics)

SETUP	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
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

In linguistics, ChatGPT has a higher rate of maintaining correct answers than maintaining incorrect ones or being inconsistent. In instances where the answers do not match, it is either Batch 3b that arrives at the correct answer, or that both batches are incorrect. This further illustrates ChatGPT's ability in being consistently correct. Furthermore, the "Give answer only" prompt only causes a minor, almost negligible, decline in ChatGPT's consistency. This further demonstrates that the model's pattern recognition abilities remain stable in linguistics questions even with addition of a prompt that limits its abilities to generate solutions.

Table 13. Comparison of ChatGPT's Answers in Batches 3a and 3b (Mathematics)

SETUP	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
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

On the contrary, in mathematics, ChatGPT has a higher frequency of being correct in Batch 3b than in Batch 3a in mathematics. Notably, it still has a higher frequency of being consistently correct than being consistently incorrect. Compared to linguistics, ChatGPT's tendency of being frequently more correct without the "Give answer only" prompt in mathematics could be a result of how mathematics questions go beyond pattern recognition. Such questions necessitate an understanding of the application of mathematical rules and reasoning abilities, skills which LLM models are not completely capable of.

Comparison of ChatGPT's Consistency between Linguistics and Mathematics

A summary of ChatGPT's consistency in each of the SAT subareas for each batch and its overall consistency in each domain was calculated through mean standard deviation. The values gathered were based on ChatGPT's raw accuracy score when it was fed SAT questions for each setup, batch, and domain. A lower standard deviation value indicates higher consistency. Table 14 lists the standard deviation of all setups for all batches.

Table 14. 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**

ChatGPT is generally more consistent with linguistics. Data shows that ChatGPT's scores in linguistics are closer to each other than in mathematics. Again, ChatGPT's more consistent performance in linguistics due to its probabilistic nature and natural

language processing capabilities leaning more towards generating and understanding human text input. As supported by simulations performed by Cheng and Yu (2023), ChatGPT's mathematical reasoning and arithmetic skills lag behind its language processing ability.

Conclusively, ChatGPT has shown its capability to aid high school students in linguistics, especially in tasks related to interpreting and analyzing passages. However, it is not entirely consistent in discerning context clues or words in sentences. In mathematics, it exhibits an inconsistent trend, with the model sometimes giving the correct answers to some questions and at other times giving different incorrect answers. Furthermore, it shows no particular strength towards any of the SAT mathematics subareas.

4.3 Survey Results

A survey among students is conducted to collect their perspectives on the objective accuracy of ChatGPT in mathematics and linguistics in relation to its potential as a student learning tool. 53 DLSU SHS Manila Grade 12 students participated, 27 of them come from the Science, Technology, Engineering, and Mathematics Strand (STEM), 15 came from the Accountancy, Business, and Mathematics Strand (ABM), and 11 from the Humanities and Social Sciences Strand (HUMSS).

The students are also asked to indicate the activities in which they use ChatGPT for. Most respondents used ChatGPT for proofreading and revising written works (66.0%), generating outline and content for written works (58.5%) , and reviewing for exams (47.2%).

To understand students' level of trust in ChatGPT, the respondents were asked to predict ChatGPT's performance. Students indicate the number of questions ChatGPT will get correctly given a ten-question exam for each subareas. Figure 19 shows the student predicted ChatGPT score and ChatGPT's actual simulation result for each subarea in linguistics and Figure 20 in mathematics.

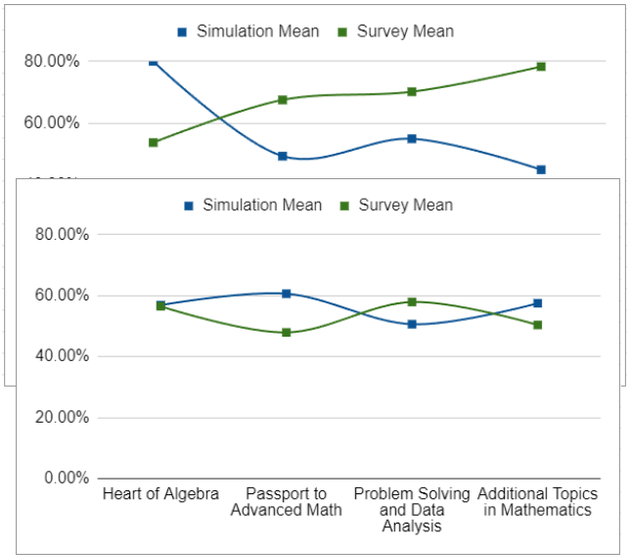

Figure 20. Comparison of Simulation Mean and Survey Mean in Mathematics

Overall, students generally believe that ChatGPT performs better than in actuality in all linguistics subareas, except for Command of Evidence. On the other hand, students believe the converse for mathematics subareas, except Problem Solving and Data Analysis. A greater gap is observed between the survey score and the simulation score in linguistics subareas over mathematics. Hence, students still need to understand ChatGPT's performance in various subjects so that they may know when and how to use this artificial intelligence tool in supplementing them in completing academic tasks.

4.4 Interview Results

Interview (Linguistics)

To understand the perceptions of professors with regards to the adoption of ChatGPT in the education system and its proficiency in Linguistics, four professors were interviewed,

whose identities are organized in Table 15 below:

Table 15. Perceptions of Professors of integrating ChatGPT in linguistics

Code	Description	ChatGPT Activities related to linguistics
L0	She is from the Department of English and Applied Linguistics in De La Salle University, where she is given the chance to teach Senior High School students, college students, and sometimes MA students.	She uses ChatGPT to further understand its nature and how her students may use them.
L1	They are a faculty of the STEM department of the De La Salle University Integrated School. At the time of writing, they teach research but have experience in teaching General Biology 2 subjects.	They frequently use ChatGPT in linguistics, specifically to check grammar, paraphrase, or improve text and sentence structure.
L2	They have experience in teaching General Biology 1 subject in De La Salle University.	They use ChatGPT to improve quiz questions, which they improve to achieve their objectives.
M2 (answered for both linguistics and mathematics)	They are a professor in De La Salle University. They teach Chemistry, handling General Chemistry 1 and Water Science and Sustainability for students from France.	They use ChatGPT to correct their grammar.

On which Domain ChatGPT is More Reliable in: Linguistics or Mathematics?

All professors interviewed for linguistics mentioned that they believe ChatGPT is more reliable in linguistics over mathematics because of prior negative and a lack of experience in using ChatGPT in mathematics, prior positive experience in using ChatGPT in linguistics-related tasks, and the design of ChatGPT to be more linguistically inclined.

Ratings

To quantify how much the perceptions of the professors interviewed aligned with the actual results of ChatGPT in each linguistics subarea during the simulations, the professors were asked to predict how many questions ChatGPT would get correctly if it were fed ten SAT questions from a certain subarea. The results are organized in Table 16 below:

Table 16: Interviewees' Predictions Versus Actual Scores of ChatGPT in Linguistics SAT Subareas

	Predictions (/10)			
Code	Standard English Conventions	Expression of Ideas	Relevant Words in Context	Command of Evidence
EB	8	5	8	5
L1	4	5	6	6
L2	7	8	7	7
M2	10	10	10	9
Mean	**7.25**	**7.00**	**7.75**	**6.75**
True Score (/10)	**4.50**	**5.50**	**4.94**	**8.00**
Percent Difference	**61.10%**	**27.27%**	**56.88%**	**15.63%**

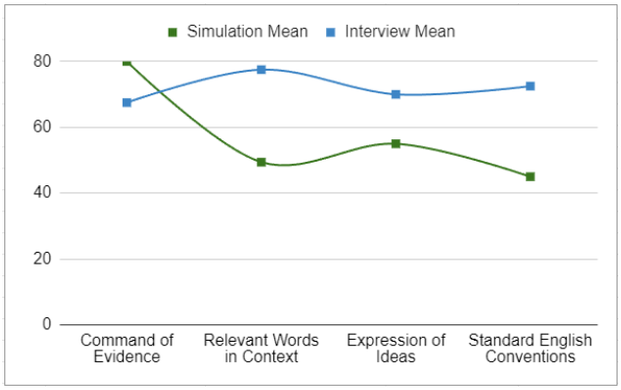

Figure 21. Comparison of Interview Mean and Simulation Mean in Linguistics

Similarly to Figure 19, Figure 21 highlights a relatively large gap between the mean of the predictions of the professors and ChatGPT's actual scores in the linguistics portion of the SATs. This supports the need for more discourse towards understanding ChatGPT's abilities for more informed decision making on its usage. The following observations were also noted:

1. The professors seem to give scores in the middle of the zero to ten spectrum, except for M2 who gave extreme prediction scores of mostly ten.
2. In most subareas, the professors seem to overestimate the proficiency of ChatGPT in linguistics by two to three points. This may be attributed to the

design of ChatGPT to be a language model in itself, leading to higher expectations. However, in the case of Command of Evidence, the professors generally underestimated ChatGPT's capabilities. When asked to explain why, they mentioned past experiences and ChatGPT's inability to include citations.

3. Interestingly, the subarea the professors believe to be ChatGPT's greatest relative weakness was the subarea ChatGPT achieved the highest score in.

Areas for Improvement

L0 and other professors have acknowledged that ChatGPT would need to improve on including in-text citations in the content it generates, especially if it got the information from a specific source.

Interview (Mathematics)

To understand the perceptions of professors with regards to the adoption of ChatGPT in the education system and its proficiency in Mathematics, four professors were interviewed, whose identities are organized in Table 17 below:

Table 17: Perceptions of Professors of integrating ChatGPT in mathematics

Code	Description	ChatGPT Activities related to Math
M0	He is a professor under the software technology department of the College of Computer Science of the De La Salle University. He teaches artificial intelligence and machine learning, and these computer technologies, which he mentions are possible because of mathematical foundations.	He uses ChatGPT to prepare some of the materials he uses for teaching, but he always verifies what ChatGPT is saying to him.
M1	M1 taught General Mathematics, Pre-Calculus subjects, Statistics and Probability, and Business Mathematics from the De La Salle University Integrated School Senior High Department.	None
M2 (answered for both linguistics and mathematics)	M2 are a professor in De La Salle University. They teach Chemistry, handling General Chemistry 1 and Water Science and Sustainability for students from France.	None.
M3	He is from the Chemical Engineering Department under the Gokongwei College of Engineering in De La Salle University. He teaches a wide range of mathematics subjects, mainly up to advanced mathematics for the chemical engineering graduate students all the way down to General Mathematics for Senior High.	He has tried using ChatGPT for Mathematics to test its abilities.

On which Domain ChatGPT is More Reliable in: Linguistics or Mathematics?

Generally, the interviewed professors believe that ChatGPT is more reliable in linguistics over mathematics. When asked to elaborate, M3 recalls his experiences where ChatGPT was consistently unable to answer basic trigonometry questions, despite trying to train it. He reasons that ChatGPT is a "*sophisticated parrot*" that copies sentences and sentence structures from a wide variety of sources on the web. This, at least, in the perspective of M3, ChatGPT can perform very well in reading comprehension, even if it has the tendency to "*hallucinate or create facts out of nowhere*". This is a similar case with M2s reasoning, who links it to their positive experience in using ChatGPT for grammar checking as they have not used ChatGPT for mathematics.

M1, on the other hand, believes that the reason why ChatGPT is more reliable in linguistics is due to the nature of mathematics questions, where one would need to have an accurate solution. According to them, "*There might be different solutions, but only one answer,*" which is why they do not think ChatGPT could give us a very reliable answer to math questions, not unless it would not require a solution.

Relating it to his knowledge from the computer science field, M0 claimed that ChatGPT should perform generally better in linguistics than mathematics because of the way it is designed to work as a conversational language model and not a logic model. According to him, "*In order to solve mathematical problems, you need some kind of logic model that understands how the different numbers relate to each other, and ChatGPT simply does not have that. And linguistics questions are all about grammar, and grammar is already captured in the word orderings and the statistical probabilities of how different words appear after one another as opposed to math.*"

Ratings

To quantify how much the perceptions of the professors interviewed aligned with the actual results of ChatGPT in each mathematics subarea during the simulations, the professors were asked to predict how many questions ChatGPT would get correctly if it were fed ten SAT questions from a certain subarea. The results are organized in Table 18 below:

Table 18: Interviewees Predictions Versus Actual Scores of ChatGPT in Mathematics SAT Subareas

	Predictions (/10)			
Code	**Heart of Algebra**	**Problem Solving and Data Analysis**	**Passport to Advanced Math**	**Additional Topics in Mathematics**
M0	7	9	6	9
M1	7	8	7	7
M2	10	10	10	10
M3	0	0	0	0
Mean	**6.00**	**6.75**	**5.75**	**6.50**
True Score (/10)	**5.69**	**5.06**	**6.06**	**5.75**
Percent Difference	**5.45%**	**33.40%**	**5.12%**	**13.04%**

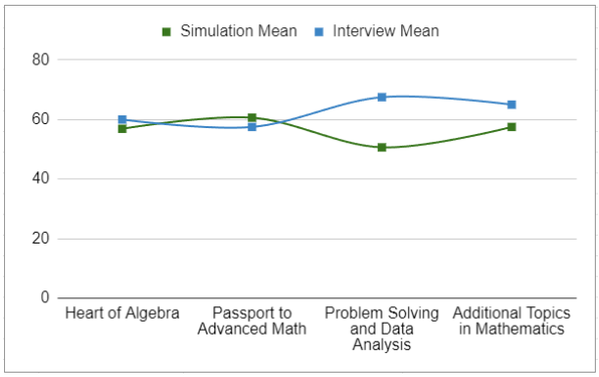

Figure 22. Comparison of Interview Mean and Simulation Mean in Mathematics

In Figure 22, the gap between the interview mean and the simulation mean are more minimal compared to linguistics. This indicates that the perception of the professors in ChatGPT mathematical abilities in the mathematics SAT subareas are similar to its actual performance. This is except for the sub area of Problem Solving and Data Analysis, where larger gaps are observed respectively. The following observations were noted:

1. The interviewed professors have varied opinions on ChatGPT's reliability, with two of them giving extreme predictions of 0 and 10 (M3 and M2). The other two interviewees gave predictions greater than 5,

indicating that to a certain extent, they believe ChatGPT would get more questions correctly than incorrectly.

2. In relation to the actual result in Heart of Algebra, Problem Solving and Data Analysis, and Additional Topics in mathematics— the professors generally gave predicted scores greater than the actual score achieved by ChatGPT. The results also show that the professors relatively underestimate the abilities of ChatGPT in Passport to Advanced Mathematics, and by extension, ChatGPT's abilities in expressing mathematical structures, reasoning with more complex equations, and building functions.

3. However, noting that the point differences of the professors' predictions and ChatGPT's actual score only ranges from 0 to 2, it can be surmised that generally, their beliefs are aligned with ChatGPT's actual proficiency.

4. Interestingly, the subarea the professors believed ChatGPT would perform most proficient in is the subarea it actually performed least accurately in (Problem Solving and Data Analysis). The converse applies, where the subarea with the lowest mean predicted score is the subarea with the highest actual score (Passport to Advanced Math).

Areas for Improvement

When asked about possible improvements on the current state of ChatGPT such that it is able to answer mathematical questions more proficiently, M3 suggests collaborations with WolframAlpha since their weaknesses are the strengths of ChatGPT and vice versa. By integrating the two models, he believes that ChatGPT can become a more reliable tool for students in mathematics-related tasks.

On the other hand, M0 talks about how ChatGPT does not have a logic model, which is why one cannot expect it to be able to answer every kind of mathematical model because it is not designed that way. However, if one were to develop ChatGPT to support the feature of solving mathematics problems, ChatGPT should be integrated with some language model. However, in computer science, ChatGPT has a problem of interpretability, which means that even if ChatGPT provides an answer, it is not able to properly explain how it arrived at that answer. In this case, it does not know that it is solving a SAT Math problem. It is simply trying to find the correct words to appear after a thread of other words. According to M0, "*This is an open problem (still) in computer science, so as of this moment, there is still a lot of research trying to solve this interpretability problem. But so far, we have no good solutions yet*."

Should ChatGPT be Integrated into the Education System?

M0 and L2 asserted that the question should not be about whether ChatGPT should be integrated into the education system; rather, it should focus more on how the education system should adapt to the emergence of ChatGPT as it is something that no one can stop. According to him, "*ChatGPT is an evolution of technology, the same way the internet is an evolution in technology*,"especially considering "*people are seeking to improve this type of technology*" as L2 puts it. Therefore, instead of trying to ban it because of concerns and apprehensions raised by academic stakeholders on this kind of technology, efforts should focus either on evolving the education system to work with this AI (M0) or placing proper guidelines on its usage within academic institutions (L2).

One way to do so, according to M0, is to design the pedagogy and curriculum such that more emphasis is placed on human skills rather than memorization of facts that can be

online. For example, assessments can be more like ' *How can you apply the things you have learned in the classroom to solve problems that are personal to you or that are personalized to your local community?'* which ChatGPT may not be able to answer.

The other mathematics professors seem to align with M0'a perspective. For instance, M3 mentioned that while he would personally not use it as a professor for crafting test questions, it can be used to suggest better ways to make them less vague. M1 supplements this by raising that ChatGPT can be used to counter check existing academic materials for potential improvements.

In relation to linguistics, M2 also agreed that ChatGPT should be integrated into the education system with limitations. While they commended ChatGPT as the "*best companion*" for grammar and that it is good for language translation, they emphasized that ChatGPT should only be used for those purposes and not for idea generation.

Of course, some professors have also mentioned the need for proper guidelines should ChatGPT be formally integrated into the education system. For example, L1 highlighted the need for more seminars and conferences for both students and professors to better identify what is ethical and unethical in using ChatGPT. This is what L0 has been trying to do in her class– to teach her students on how to properly use ChatGPT for academic purposes.

Nevertheless, to summarize, most professors seem to agree that in one way or another, ChatGPT should be integrated into the education system.

Is ChatGPT a Reliable Learning Tool for Students?

M3 strongly says no, considering the inability of ChatGPT to accurately and consistently answer mathematics questions. However, he accepts that ChatGPT can be used to help students express their ideas more clearly.

S2, who also expressed that ChatGPT is not a reliable learning tool, added that students are not disciplined enough to use this kind of AI, particularly because they end up not using it for learning purposes (e.g. to generate content, to get high grades, etc.), which defeats the purpose of having an education. For them, ChatGPT can only be a reliable tool if there are guidelines that students can follow to ensure proper use of the software.

Conversely, M1 and S1 say yes with consideration that students should use it responsibly, which includes avoiding academic dishonesty (M1) and fact checking (S1). Whatever information from ChatGPT should be backed with reliable studies. M0 adds, "*ChatGPT should be integrated into the learning process. You can use it as a guide to explore further into the topics you are currently studying. It also enables you to search for answers quickly and in a more natural way, so it really eases up things that would otherwise be tedious if you don't have it."*

While M2 also agreed that ChatGPT is a reliable tool, they do not recommend using the chatbot for technical subjects, such as science, physics, and mathematics. He surmised that ChatGPT is only best for linguistics. L0 supplements this by highlighting her 50-50 rules wherein students can only rely on ChatGPT 50% percent of the time, and 50% of the time, students should stay true to their ideas. There has to be some form of balance, even if there seems to be a consensus that ChatGPT has exhibited potential to be a reliable learning tool for students, assuming it is used properly.

CONCLUSION

Recently, ChatGPT has been popularized as an aid for students to learn due to its easy accessibility and ability to converse using

simple language. Based on the experiments conducted, ChatGPT's performance outperforms the average high school student test taker's performance in both linguistics and mathematics, hinting at its potential to assist students. However, it was observed that ChatGPT struggles with context-related tasks in linguistics and in advanced mathematics concepts. Hence, the use of ChatGPT should be with caution, especially for tasks relevant to the aforementioned.

Despite a slight edge in linguistic accuracy, there's a gap between actual and perceived scores, highlighting the need for clear understanding of ChatGPT's capabilities. Professors acknowledge its potential in linguistics but stress the importance of having a set of guidelines on its use. Nevertheless, they see it as a valuable learning aid for improving student expression.

Moreover, prompting has shown to affect the performance of ChatGPT. This means that prompts may be used to manipulate ChatGPT's accuracy and consistency. Users must be careful in using prompts as these prompts must be selected carefully.

Lastly, the study highlights a need for proper guidelines regarding the use of ChatGPT, especially one that is backed by research such as this one. Further studies may delve into understanding factors affecting ChatGPT's performance.

FUTURE WORK

Future research can tackle ChatGPT's performance in answering questions in other subject domains, such as science and history, to determine the chatbot's effectiveness as a learning tool in these academics. Additionally, this can also guide AI developers in pinpointing the certain areas, topics, and subjects where current AI tools are inefficient in and how much their users should rely on the outputs of these tools. Future works can also determine ChatGPT's performance in specific areas under the subject domains (ex. arithmetic expressions, geometrical figure analysis, and statistical interpretation under mathematics).

Considering that this is one of limited studies tackling the consistency of ChatGPT's answers, future research can also focus on further understanding and quantifying these consistencies and inconsistencies. This includes determining the factors that could affect ChatGPT's consistency, and overall performance, in answering questions. Doing so would aid high school students in understanding ChatGPT's capabilities as a learning tool.

As prompts have been established to significantly impact ChatGPT's performance, future works may also consider looking into the effect of prompt engineering techniques such as few-shot learning. Few-shot prompting involves training the LLM on what it needs to do by providing examples within the initial prompt. Fong and Ong (2024) have already demonstrated how giving ChatGPT examples via few-shot learning can improve its performance on joint Intent Detection and Slot-Filling tasks, which involve identifying the purpose and key slots (information) behind a user's query. Other studies that seek to optimize ChatGPT's performance may also focus on the construction and identification of the prompts required for ChatGPT to respond to a question with the highest levels of accuracy and consistency. The formation of such prompts may aid ChatGPT in achieving an ideal and satisfactory performance.

REFERENCES

BakerMcKenzie. (2020). *So you think you want to. . .use AI*. https://www.bakermckenzie.com/-/media/restricted/cgr/so-you-think-you-want-to-use-ai.pdf

Barros, C. (2023). ChatGPT in the resolution of a math exam: Results obtained in Portuguese and in English language. Proceedings of the International Conference on Lifelong Education and Leadership for All (ICLEL 2023) (pp. 37-47). Atlantis Highlights in Social Sciences, Education and Humanities. https://doi.org/10.2991/978-94-6463-380-1_5

Bogost, I. (2023). Is this the singularity for standardized tests? *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/03/open-ai-gpt4-standardized-tests-sat-ap-exams/673458/

Bonsu, E. M., & Baffour-Koduah, D. (2023). From the consumer's side: Determining students' perception and intention to use ChatGPT in Ghanaian higher education. *Journal of Education, Society, & Multiculturalism, 4*(1), 1-29. https://doi.org/10.2478/jesm-2023-0001

Cheng, V., & Yu, Z. (2023). Analyzing ChatGPT's mathematical deficiencies: Insights and contributions. The 35th Conference on Computational Linguistics and Speech Processing (pp. 188-193). ACL Anthology. https://aclanthology.org/2023.rocling-1.22/

Choi, J.H., Hickman, H.E., Monahan, A., & Schwarcz, D. (2023). ChatGPT goes to law school. *Social Science Research Network Electronic Journal*. http://dx.doi.org/10.2139/ssrn.4335905

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv.org. https://doi.org/10.48550/arXiv.2110.14168

College Board. (2023). SAT program results for the class of 2023 show continued growth in SAT participation. *Newsroom | College Board*. https://newsroom.collegeboard.org/sat-program-results-class-2023-show-continued-growth-sat-participation

Crabtree, M. (2023). What is ChatGPT? A chat with ChatGPT on the method behind the bot. DataCamp. https://www.datacamp.com/blog/a-chat-with-chatgpt-on-the-method-behind-the-bot

De Winter, J. (2023). Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education*. https://www.researchgate.net/publication/366659237_Can_ChatGPT_pass_high_school_exams_on_English_Language_Comprehension

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International, 60*(1), 1-15. https://doi.org/10.1080/14703297.2023.2195846

Firat, M. (2023). What ChatGPT means for universities: Perceptions of scholars and students. *Journal of Applied Learning & Teaching, 6*(1), 57-63. https://doi.org/10.37074/jalt.2023.6.1.22

Frieder, S., Pinchetti, L., Griffiths, R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). Mathematical capabilities of ChatGPT. *ResearchGate*. https://doi.org/10.48550/arxiv.2301.13867

Fuchs, K. (2023). Exploring the opportunities and challenges of NLP models in higher education: Is Chat GPT a blessing or a curse? *Frontiers in Education*, 8. https://doi.org/10.3389/feduc.2023.1166682

Fong, H., & Ong, E. (2024). Evaluating ChatGPT for joint intent detection and slot filling: Zero-shot vs. few-shot prompting. In *Proceedings of Philippine Computing Science Congress 2024*. Computing Society of the Philippines.

Korkmaz, A., Aktürk, C., & Talan, T. (2023). Analyzing the user's sentiments of ChatGPT using Twitter data. *Iraqi Journal for Computer Science and Mathematics, 4*(2), 202-214. http://dx.doi.org/10.52866/ijcsm.2023.02.02.018

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. Humanities and Social Sciences Communications, 10, 704. https://www.nature.com/articles/s41599-023-02119-6

Li, C., & Xing, W. (2021). Natural language generation using deep learning to support

MOOC learners. *International Journal of Artificial Intelligence in Education, 31*(2), 186-214. https://doi.org/10.1007/s40593-020-00235-x

Mohamed, A. (2023). Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a foreign language (EFL) teaching: Perceptions of EFL faculty members. *Education and Information Technologies*. https://link.springer.com/article/10.1007/s10639-023-11917-z

Muehmel, K. (2023). What is a large language model, the tech behind ChatGPT? Dataiku. https://blog.dataiku.com/large-language-model-chatgpt

Muniz, H. (2021). The 4 SAT sections: What they test and how to do well. *PrepScholar*. https://blog.prepscholar.com/sat-sections

Nazario, M.E.K., Tan, E.J., Lim, S.V., Ramos, K.A., & Chu, S. (2024). AI skill check: An examination of ChatGPT's Consistency in Linguistics and Mathematics using the SAT. In *Proceedings of Philippine Computing Science Congress 2024*. Computing Society of the Philippines.

Ohsie-Frauenhofer, E. (2023). The SAT will become fully digital - and shorter - by 2024. Here's what's changing and what's staying the same. *ArborBridge*. https://blog.arborbridge.com/sat-will-become-fully-digital-and-shorter-by-2024-whats-changing

Rogerson, K. (2023, March 1). Higher Ed Beware: 10 Dangers of ChatGPT in Education that Schools Need to Know. Comm100. https://www.comm100.com/blog/higher-ed-beware-10-dangers-chatgpt/

Shakarian, P., Koyyalamudi, A., Ngu, N., Mareedu, L. (2023). An independent evaluation of ChatGPT on mathematical word problems (MWP). *arXiv.org*. https://doi.org/10.48550/arXiv.2302.13814

Tan, E.J., Ramos, K.A., Nazario, M.E.K., Lim, S.V., & Chu, S. (2024). AI to the test: Measuring ChatGPT's objective accuracy in answering the SATs in comparison to human performance. In *48th IEEE International Conference on Computers, Software, and Applications*.

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

USAFacts Team. (2022). *Are fewer students taking the SAT?* https://usafacts.org/articles/are-fewer-students-taking-the-sat/

Wardat, Y., Tashtoush, M., Alali, R., & Jarrah, A. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *EURASIA Journal of Mathematics, Science, and Technology Education, 19*(7). https://doi.org/10.29333/ejmste/13272

Wolfram, S. (2023). What is ChatGPT doing… and why does it work? Stephen Wolfram Writings. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

Yavuz, Z. (2024). Why is ChatGPT bad at even basic math? Retable. https://www.retable.io/blog/why-is-chatgpt-bad-at-math

Zvornicanin, E. (2024). Why is ChatGPT bad at math? Baeldung on CS. https://www.baeldung.com/cs/chatgpt-math-problems