ChatGPT’s Performance in Spreadsheets Modeling Assessments based on Revised Bloom’s Taxonomy

Michelle CHEONG
School of Computing & Information Systems, Singapore Management University
michcheong@smu.edu.sg

Abstract: ChatGPT has taken the education scene by storm and caused uneasiness among educators. Mixed reactions were observed with some institutions banning it, while others embracing it with caution. This paper evaluates the performance of ChatGPT on solving spreadsheets modeling assessment questions with multiple test items categorized according to the revised Bloom’s taxonomy, to discover the accuracy of the answers provided at each cognitive learning level. The insights obtained may be useful for educators to design future assessment questions which focus more on testing critical thinking skills to assess the students accordingly to achieve the intended learning outcomes, and we propose recommended actions on how to do so. Our proposed methodology can be applied to other course modules to achieve their respective insights for future assessment designs and actions.

Keywords: ChatGPT, Performance Evaluation, Spreadsheets Modeling, Assessments, Revised Bloom’s Taxonomy

1. Introduction

ChatGPT (Generative Pre-trained Transformer) is a Generative AI (GAI) Large Language Model (LLM) developed by OpenAI (OpenAI, 2022) trained to provide answers across a myriad of domains. Released in November 2022 as free access to all users, New York City (The Guardian, 2023) and New South Wales and Queensland in Australia (ABC News, 2023) decided to ban it in schools with many other cities following suit. Such a reaction is reasonable as CNN reported that ChatGPT was able to pass the law exam and a business school exam (CNN Business, 2023). However, when ChatGPT was tested on three years of Singapore’s Primary School Leaving Examinations (PSLE) for 6th grader, it failed miserably in Mathematics and Science, and only managed a borderline pass for English (Business Insider, 2023). Many papers supported using ChatGPT for teaching and learning, while at the same time expressed many concerns (Alkaissi & McFarlane, 2023; Gilson et al., 2023; Kung et al., 2023). Thus, it is not clear to many whether ChatGPT, or any other GAI like Google Bard or Amazon Bedrock, is really a boon or bane to education.

In this paper, we review the recent literature on ChatGPT’s implications on education and how it performed in terms of educational assessments. We then evaluate ChatGPT’s performance in quiz questions for a university spreadsheets modeling module where the questions with multiple test items are mapped according to the revised Bloom’s taxonomy. With the answers provided by ChatGPT, we establish how well ChatGPT tackled technical questions on the different cognitive learning levels. We hope to provide educators with insights to design future assessments on spreadsheets modeling module and propose recommended actions on how to do so, especially for institutions which are ready to embrace ChatGPT in education. Our proposed methodology can be applied to other course modules to achieve their respective insights for future assessment designs and actions.

2. Literature Review

Due to the recent rapid increase in the popularity of ChatGPT, several academic papers were published on its implications on education and its performance in educational assessments. Lim et al. (2023) presented four paradoxes of GAI and their implications to the future of education. “Paradox 1: GAI is a friend yet a foe” highlighted its advantages in accelerated learning, discovery of new knowledge, reform the design of assessments to achieve higher level, and yet concerned with misinformation and its impact on academic integrity. “Paradox 2: GAI is capable yet dependent” discussed the need to make the best use of it to maximize returns, and yet be cognizant that it is dependent on the quantity and quality of prompts and its prior training. “Paradox 3: GAI is accessible yet restrictive” discussed the need for a sustainable model to promote equitable access to all. “Paradox 4: GAI is popular when banned” promoted its use in education rather than banning it which may lead to many negative consequences. Through the four paradoxes, they provided guidelines for practices and directions for future research.

Pavlik who co-authored with ChatGPT (2023) studied its implications on journalism and media. ChatGPT could provide high-quality answers and thus may pose a threat to journalists and media professionals. He suggested educators to use it to develop course but warned that students might use it in their own academic work compromising academic integrity. Alkaissi and McFarlane (2023) presented two medical cases and asked ChatGPT to provide scientific writing on specific medical and non-medical topics. The response provided was mostly factually correct but with some errors. The references provided were non-existent, demonstrating that it suffered from “artificial hallucination” where seemingly realistic outputs do not correspond to any real-world inputs. Thus, they advocated using AI output detectors in editorial processes and clear disclosures by authors. Similarly, Anderson et al. (2023) also requested ChatGPT to write two academic papers on using AI for scientific manuscripts and in sports medicine. They also found that the bibliographies generated were wrong. One concerning finding was that when using GPT-2 Output Detector to assess the originality score for both papers, the scores increased tremendously by simple paraphrasing.

In terms of ChatGPT’s performance in educational assessments, Gilson et al. (2023) and Kung et al. (2023) both evaluated ChatGPT’s performance in the United States Medical Licensing Exam (USMLE). Gilson et al. (2023) used questions from question banks for Step 1 and Step 2 exams, while Kung et al. (2023) selected sample exam questions from June 2022 to test ChatGPT’s performance on answering original questions instead of retrieving answers from the corpus which it was trained on. Both studies concluded that ChatGPT passed the threshold of 60% and can potentially assist learners in medical education. One interesting finding from Gilson et al. (2023) was that the accuracy of the responses significantly decreased as the difficulty of the question increased. They measured the types of errors committed by ChatGPT and found that it committed mostly logical errors (failure to convert information to answer), then information errors (failure to identify key information), and very few statistical errors (arithmetic error).

With ChatGPT performing reasonably well in educational assessments, how should future assessments be changed? Stokel-Walker (2022) suggested the possibility of removing essay writing altogether. Zhai (2022) suggested educators to focus on improving students’ creative and critical thinking skills instead of general skills, and to use AI tools to conduct subject-domain tasks. Echoing this, O’Connor who co-authored with ChatGPT (2023) recommended to use oral presentation and objective structured clinical examinations to increase the diversity of assessments used in nursing education.

In this paper, we evaluate the accuracy performance of ChatGPT to solve technical questions on spreadsheets modeling and each question with multiple test items was mapped against the revised Bloom’s taxonomy. Our work is different from previous works in two main areas. Firstly, past works mostly evaluated ChatGPT’s performance in non-technical modules, while we used a technical module in spreadsheets modeling. The only other work that used a technical module is the unpublished work by Malinka et al. (n.d.) where they evaluated ChatGPT’s performance in computer security modules. They only shared the types of assessments used to test ChatGPT but did not share any details of the questions.

They found great variability in the correctness of ChatGPT's answers and artificial hallucination also occurred. Secondly, none of the previous works linked ChatGPT's performance on test items categorized on a cognitive learning scale, while we mapped test item against the well-accepted revised Bloom's taxonomy to achieve insights to advise future assessment designs in response to recommendations by Zhai (2022) and Lim et al.(2023) to focus more on assessing critical thinking skills.

3. Bloom's Taxonomy and Spreadsheets Modeling Module

3.1 Bloom's Taxonomy

Bloom's taxonomy is a theoretical model (Bloom, 1956) intended to classify cognitive learning from simple to complex level in a progressive manner, to aid educators in reducing duplicate assessment items to achieve the same learning outcome. There are six levels (1) knowledge, (2) comprehension, (3) application, (4) analysis, (5) synthesis, and (6) evaluation. For each level, there are action verbs to classify assessments accordingly. To represent the 21st century teaching and learning, Anderson and Krathwohl (2001) revised Bloom's taxonomy into a two-dimensional model. The six levels remained but they were changed from noun to verb as (1) remember, (2) understand, (3) apply, (4) analyze, (5) evaluate, and (6) create. Levels 5 and 6 were swapped, and a second dimension called the “knowledge dimension” which has four levels (1) factual, (2) conceptual, (3) procedural, and (4) metacognitive, was added in addition to the original “cognitive dimension”.

One main challenge in applying Bloom's taxonomy stems from the overlapping action verbs in multiple levels, leading to ambiguity in its application as highlighted by Das, Mandal, and Basu (2022). Even with the revised taxonomy, the ambiguity issue did not go away totally. In this paper, we classify our test items to the different levels of the revised Bloom's taxonomy based on two module instructors' expertise in the subject matter, and a third instructor was consulted in case of different classification.

3.2 Spreadsheets Modeling Module

We use two quizzes created for a university level module in spreadsheets modeling to evaluate ChatGPT's performance. The module objective is to teach students how to build spreadsheet models from scratch and perform calculations and data analysis to support business decision making. The learning outcomes include formulating business problems and integrating business analysis skills to model and appraise business problems; acquiring computer skills to perform problem analysis; and acquiring competency in spreadsheets tool. The module covers six main topics as described in Table 1.

Table 1. *Six Main Topics Covered in the Spreadsheets Modeling Module*

Topic	Description
1	Basic modeling techniques to formulate arithmetic equations and using Excel functions to perform basic statistical calculations.
2	Spreadsheet engineering skills in terms of using conditions to evaluate outcomes, create basic charts, compute intercept and slope of linear lines, create different order of polynomial functions to represent business logic relationships to be included into the model to seek for best answer using Goal Seek.
3	Financial calculations using Excel functions to understand and compute time value of money to assess investment options and returns.
4	Data lookup and linkup to handle changing inputs which affect outputs, and to determine optimal solutions to business problems which are subjected to constraints using Solver Add-in.

5	Monte-Carlo simulation to simulate repeated scenarios using simulated input values from re-sampling, cumulative relative frequency, and distribution functions, to obtain repeated output results using Data Table, to support decision making.
6	Extends topic 5 to include date and time for time-based discrete event simulation to determine performance of queue systems.

3.3 Assessment Quizzes

The module instructor designed the two quizzes for the spring term of academic year 2022/23. As the quiz questions were original, they did not exist before January 2023, and thus will not form part of the ChatGPT's training corpus. Each quiz describes a business scenario and contains multiple questions as listed in Table 2. Q1 to Q5 are for Quiz 1, while Q6 to Q10 for Quiz 2. Each question contains multiple test items, and each item is mapped to the revised Bloom's taxonomy (BT), with the lowest at 3, and the highest at 6. It is not common to test students on levels 1 and 2 for university level modules. The mapping was done by two instructors and the inter-rater agreement between them was evaluated using the Cohen's kappa statistics (Hallgren, 2012), and found to be 0.902.

Table 2. *Questions and Test Items from Two Quizzes and their Bloom's Taxonomy (BT)*

Question	Test item	BT
Q1. A mall offers different discounts, Types A to F, and allows shoppers to split their total purchase amount into multiple payments to enjoy as many discounts as possible. Type A is a \$10 discount with minimum \$140 spending. Type B is a \$20 discount with a minimum \$500 spending. Type C is a \$10 discount with a minimum of \$125 spending. Type D is a \$45 discount with a minimum spending of \$800. Type E is a \$15 discount with a minimum \$800 spending. Type F is a 1.5% discount with no minimum spending. Compute the ratio defined as discount divided by minimum spending, and compute the rank of the ratio in descending order using a suitable Excel function. Create an Excel spreadsheet model, called Table 1, to determine the answer.	• Test arithmetic to compute ratio. • Test identification and application of correct Excel function to rank ratio.	3 4
Q2. You plan to buy a computer for \$1200 and wish to split the purchase amount into multiple payments to use as many discounts as possible, to enjoy the maximum discount. Create a second Excel spreadsheet model, called Table 2. In the first column, list the rank from 1 to 6. Then use any Excel lookup function to retrieve the discount types from Table 1 corresponding to the rank position, displaying the discount, the amount to be charged to this discount, and the discount enjoyed. It is only smart to charge the minimum amount to	• Test identification and application of lookup function to retrieve data. • Test formulation of complex formulas to determine if discount is used and amount charged to each discount considering all	3 6

the discount if this discount is used. Then add another column to indicate if the discount is used using "Yes" or "No" label. Finally, compute the amount payable if a discount is used, and the final total amount payable.	discounts collectively to enjoy maximum discount. • Test amount payable if discount is used. • Test computation of total amount payable.	 5 3
Q3. With the final total amount computed earlier, you can choose to pay it over three equal monthly instalments. Using an annual interest rate of 2%, what would be the present value of the final total amount?	• Test identification and application of correct Excel function to perform financial calculation.	4
Q4. Your friend wants to buy your computer for $1400 today. How much would you earn today if you were to sell your computer to your friend, based on the present value of the final total amount you paid computed earlier?	• Test arithmetic to compute earnings.	3
Q5. At what annual interest rate would you be able to earn $300 today if you were to sell your computer to your friend for $1400, based on the present value of the final total amount you paid using this new annual interest rate?	• Test recursive calculation using Goal Seek or Solver.	5
Q6. At a clinic, patients join a single queue lining up one after another to see the doctor. Assuming the inter-arrival time of patients follows an exponential distribution with mean 4 minutes. Out of these patients, 35% of them see the doctor due to suspected COVID-19 infection, while the remaining 65% are normal patients. For those who suspect COVID-19 infection, the doctor will test to confirm if they are indeed infected. It was found that 70% of them are confirmed. Due to different medical conditions, the consultation time differs, and they all follow normal distributions with different means and standard deviations. For the normal patients, the mean is 5 minutes and standard deviation is 1 minute. For the suspected and not confirmed patients, the mean is 7 minutes and standard deviation is 2 minutes. For the suspected and confirmed patients, the mean is 15 minutes and standard deviation is 3 minutes. Create an Excel model Table 1 with the first column to simulate the	• Test data simulation using exponential distribution for inter-arrival time. • Test conditional probability to simulate patient type. • Test time calculation. • Test IF condition to get correct inputs for simulating consultation time using normal distribution for different patient type.	4 4 3 5

inter-arrival time of 50 patients, then compute the arrival time of the patients using 9am as the reference starting time in the second column. Then add a new column to simulate if this patient is suspected COVID-19 case using “Yes” or “No” label. If “Yes”, then add another column to simulate if this patient is a confirmed COVID-19 case using “Yes” or “No” label. Finally, add another column to simulate the consultation time depending on the patient type.		
Q7. Continuing with Table 1 created earlier, for each patient determine the consultation start time, consultation end time, the wait time, and the system time. Update as Table 2.	• Test queue system concepts. • Test arithmetic to compute start time, end time, wait time system time.	4 4
Q8. While waiting in the line, it is possible that a non-COVID-19 patient (e.g. Person B) may get infected by the person standing directly in front (e.g. Person A), and/or the person standing directly behind (e.g. Person C), if Person A and/or Person C are confirmed COVID-19 cases, and the overlap time they spend standing in line next to Person B exceeds 30 mins. Continuing with the Table 2 created earlier, add two new columns to determine the overlap time between Person A and Person B, and overlap time between Person B and Person C. Finally, add a last column to indicate for each patient, will this patient get infected using “Yes” or “No” label. However, if this patient is already a confirmed COVID-19 patient, indicate “C+” instead.	• Test queue system concepts. • Test arithmetic to compute overlap time. • Test formulation of complex formulas to determine new infection status taking into account multiple linked considerations.	4 4 6
Q9. Using the updated Table 2 created earlier, compute the number of possible new infections, and use Data Table to repeat this simulation 500 times to determine an overall average number of possible new infections.	• Test identification and application of correct Excel function to count based on test condition. • Test Data Table concept to repeat simulations. • Test arithmetic to compute average.	4 5 3
Q10. The doctor thinks that he should reduce the mean consultation time by 2	• Test Data Table concept to repeat	6

	minutes. He can choose to do that for only one of the cases, normal patients, suspected and not confirmed patients, or suspected and confirmed patients. Regardless of which case he decides to reduce the mean consultation time, the wait time for all patients will be reduced, and thus lead to reducing the number of possible new infections. Determine the new overall average number of possible new infections for each case using Data Table and conclude which case will be most effective and explain why.	simulations with different inputs to test different scenarios. • Test inference of insights from results obtained.	5

To simulate the similar answering process as real students, the questions were posted to ChatGPT sequentially in one single continuous conversation to allow it to refer to its previous responses to generate the next response for follow-up questions. This is an in-built capability that ChatGPT has, which fascinated many users.

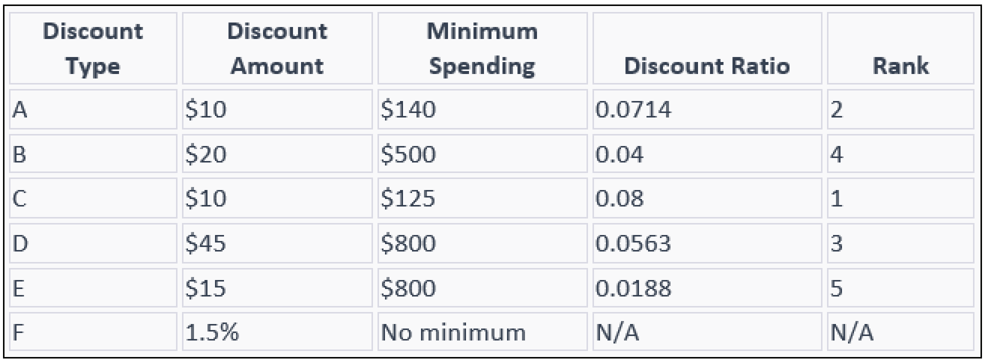
4. ChatGPT's Performance in Spreadsheets Modeling Assessment

We evaluate ChatGPT's performance for the quizzes over three complete runs and compute the average accuracy at each cognitive level. While the responses from ChatGPT were not exactly the same for each run, they were rather similar albeit different mistakes committed. The main difference comes from the way the responses were presented, which could be in the form of step-by-step explanation or presented as table. We found that ChatGPT can perform reasonably well at level 3 with 67% accuracy and the accuracy decreases as the cognitive level increases. This result is similar to the work by Gilson et al. (2023) where they reported a significant decrease in accuracy as the question difficulty increases for USMLE.

Table 3. *ChatGPT's Accuracy Performance at Different Bloom's Taxonomy (BT) Level*

BT	# items	ChatGPT Average Accuracy
3 (apply)	6	67%
4 (analyze)	9	39%
5 (evaluate)	5	33%
6 (create)	3	22%

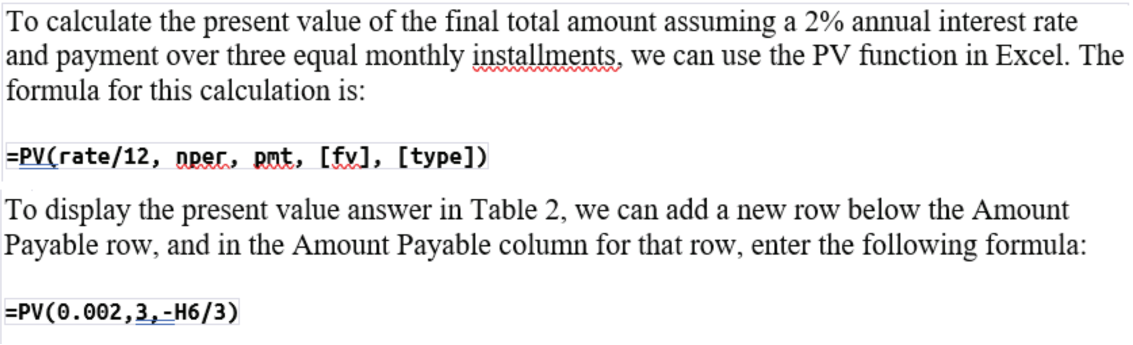
At level 3 (apply), ChatGPT was able to perform arithmetic calculations reasonably well. However, it committed mistakes when it misinterpreted the information provided in the question. For example, in Q1, the 1.5% discount for Type F is equivalent to a discount ratio of 0.015 but ChatGPT misinterpreted it and indicated it as N/A instead (Figure 1).



Discount Type	Discount Amount	Minimum Spending	Discount Ratio	Rank
A	$10	$140	0.0714	2
B	$20	$500	0.04	4
C	$10	$125	0.08	1
D	$45	$800	0.0563	3
E	$15	$800	0.0188	5
F	1.5%	No minimum	N/A	N/A

Figure 1. ChatGPT Misinterprets Information.

At level 4 (analyze), ChatGPT was able to analyze the problem and determine the correct Excel function to use but was not able to apply the function correctly. For example, in Q3, it correctly identified using PV function, but the arguments to fill in the function were wrong, where the [type] should be set as 1 and not ignored which means type 0 (Figure 2).



To calculate the present value of the final total amount assuming a 2% annual interest rate and payment over three equal monthly installments, we can use the PV function in Excel. The formula for this calculation is:

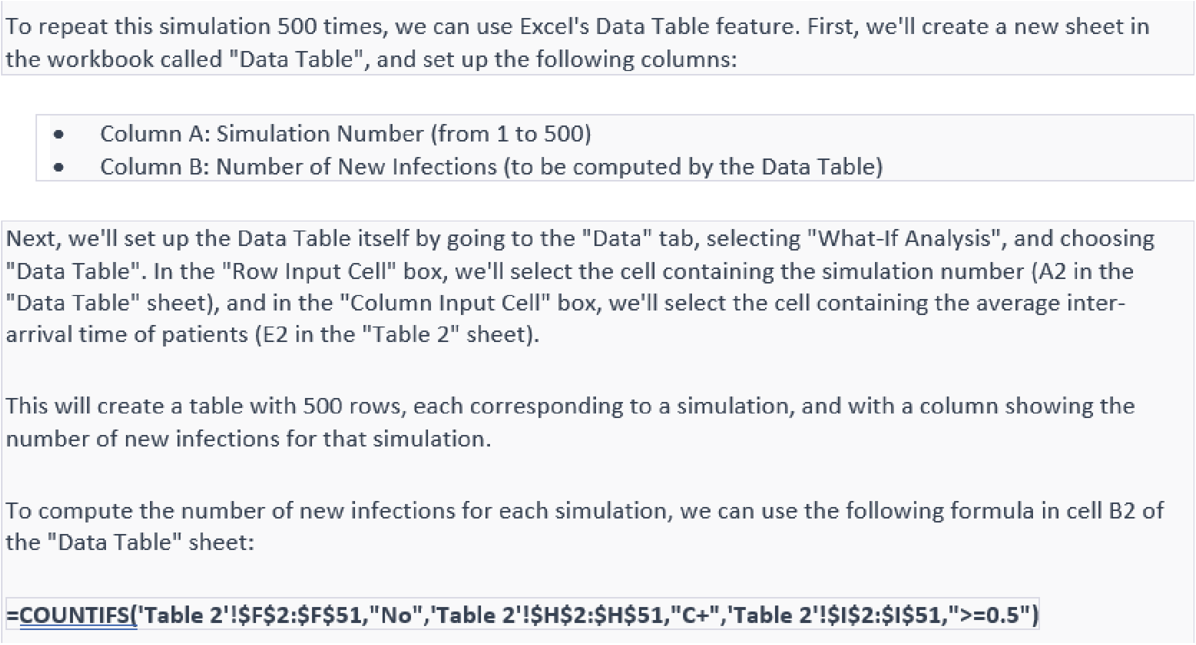
```
=PV(rate/12, nper, pmt, [fv], [type])
```

To display the present value answer in Table 2, we can add a new row below the Amount Payable row, and in the Amount Payable column for that row, enter the following formula:

```
=PV(0.002,3,-H6/3)
```

Figure 2. ChatGPT Correctly Identifies PV Function but Filled in Wrong Argument.

At level 5 (evaluate), ChatGPT was able to explain the evaluation approach, but was not able to perform the evaluation correctly. For example, in Q9, ChatGPT can explain how to use Data Table to perform the simulation, but the actual calculation using COUNTIFS function was not done correctly, as it should only count patients with new “Yes” label, and not those who are “No” and “C+” labels. (Figure 3).



To repeat this simulation 500 times, we can use Excel's Data Table feature. First, we'll create a new sheet in the workbook called "Data Table", and set up the following columns:

- Column A: Simulation Number (from 1 to 500)
- Column B: Number of New Infections (to be computed by the Data Table)

Next, we'll set up the Data Table itself by going to the "Data" tab, selecting "What-If Analysis", and choosing "Data Table". In the "Row Input Cell" box, we'll select the cell containing the simulation number (A2 in the "Data Table" sheet), and in the "Column Input Cell" box, we'll select the cell containing the average inter-arrival time of patients (E2 in the "Table 2" sheet).

This will create a table with 500 rows, each corresponding to a simulation, and with a column showing the number of new infections for that simulation.

To compute the number of new infections for each simulation, we can use the following formula in cell B2 of the "Data Table" sheet:

```
=COUNTIFS('Table 2'!$F$2:$F$51,"No",'Table 2'!$H$2:$H$51,"C+",'Table 2'!$I$2:$I$51,">=0.5")
```

Figure 3. ChatGPT can Explain Data Table Setup but Performed Wrong Calculation.

At level 6 (create), where complex formulas which involve linking several parts of calculations together are expected, ChatGPT failed miserably. It tends to over-simplify the calculations leading to the wrong answers. For example, Q2 required formulas to determine if a specific discount is used and the amount charged to each discount if that discount is used. This requires all discounts to be considered together as a whole to maximize total discount, and not treat each discount as separate consideration.

5. Insights and Implications

We found that ChatGPT's responses are often self-contradictory and incoherent when responding to technical calculation questions. It can describe the calculation process correctly, but the actual calculation was performed incorrectly. Very often, the suggested calculation steps are messy and cell references in the provided formulas are not referring to the correct cells. One plausible explanation for such a behavior is that ChatGPT is a language model. It excels in "auto-complete" and thus performed well in essay writing. However, when it comes to logical and coherent thinking required in complex calculations involving linking multiple considerations, it was not able to handle.

On the other hand, we found that ChatGPT is knowledgeable in explaining Excel functions in terms of the arguments and how to use them. And it is surprisingly good in queuing system, probability and distribution functions concepts. It could be due to the corpus it was trained on which is mainly based on textual content and facts.

Generative AI is here to stay and will become more powerful creating immense impact on education. Instead of outright banning it, educators should consider how to best use it for course development, teaching and assessment, as echoed by many researchers including Lim et al. (2023) and Zhai (2022). Based on the results of our study, we provide three recommendations for educators who teach spreadsheets modeling modules. One, allow students to use ChatGPT to answer in-class activity questions (not actual assessments), and guide them to identify errors and limitations in the answers and to suggest how to improve the answers. For example, request students to build a first model based on ChatGPT's suggested answers, and then build a second model with any identified errors corrected. Two, ask students to formulate questions to ask ChatGPT and through analyzing the answers, learn to sharpen the questions to represent the exact intent of the questions as a form of interactive learning (Rospigliosi, 2023). For example, students can ask ChatGPT to compare two different investments options, and then sharpen the question to ask for comparison using net present value or internal rate of return calculations. And lastly, design assessment questions which focus on testing higher order thinking skills, with test items from level 4 onwards (Zhai, 2022). For example, design questions that link data lookup function with conditional test to determine the answer, instead of questions that only require simple arithmetic calculations like computing the sum or average. Our recommendations aim to overcome diminishing learners' own innovative capacities and critical thinking skills, and focus on enhancing their higher order thinking skills.

6. Conclusions and Future Research

Our study shows that ChatGPT's performance in spreadsheets modeling assessment questions was good up to level 3 of the revised Bloom's taxonomy, and its accuracy decreases as the cognitive level increases. It could explain technical facts very well and performed reasonably well in basic arithmetic calculations. However, from level 4 onwards, ChatGPT was unable to apply Excel functions correctly even though it has identified the correct Excel function; unable to evaluate complex situations even though it could explain the evaluation approach; and unable to formulate complex formulas that link multiple considerations as its formulations were often simplified. However, we believe that ChatGPT and other GAI will become better and as educators we should exploit their full potential in education while keeping in mind how to overcome the concerns on cheating, plagiarism and loss of independent thinking skills in students. This calls for a balanced approach and a paradigm shift in education.

We recommend several future research possibilities involving GAI. From the instructors' perspective, we can perform research in using GAI to generate assessments and assess the cognitive levels of such assessment questions to understand if GAI can generate questions of higher order thinking skills, and how much GAI can lighten instructors' workload. From students' perspective, we can quantify the improvement in knowledge gain in students when using GAI as part and parcel of their learning process. From the GAI tool's perspective, we can consider developing and implementing customized GAI tool using LLM to teach and learn technical modules such as spreadsheets modeling or Python programming, and assess the tool's effectiveness.

References

ABC News. (2023). Queensland to join NSW in banning access to ChatGPT in state schools. ABC News. https://www.abc.net.au/news/2023-01-23/ queensland-to-join-nsw-in-banning-access-to/101884288

Alkaissi, H., & McFarlane, S.I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* 15(2): e35179. DOI:10.7759/cureus.35179

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Anderson, N., Belavy, D.L., Perle, S.M., Hendricks, S., Hespanhol, L., Verhagen, E., & Memon, A.R. (2023). AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation. BMJ Open Sport & Exercise Medicine, 9:e001568. DOI:10.1136/bmjsem-2023-001568

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. *Handbook I: Cognitive domain*. New York, NY: David McKay Company.

Business Insider (2023). ChatGPT failed miserably in Singapore's 6th-grade tests, averaging 16% for math and 21% for science. Days later, it was getting answers right. https://www.businessinsider.com/chatgpt-failed-singapore-sixth-grade-exams-psle-2023-2

CNN Business (2023). ChatGPT passes exams from law and business schools. https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html

Das, S., Mandal, S.K.D., & Basu, A. (2022). Classification of action verbs of Bloom's Taxonomy cognitive domain: An Empirical Study. *Journal of Education*, 202(4), 554-566. doi.org/10.1177/00220574211002199

Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A., & Chartash, D. (2023). How does ChatGPT perform on the United States Medical Licensing Examination? The implications of Large Language Models for medical education and knowledge assessment. *JMIR Medical Education*, 9.

Hallgren, K.A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), pp. 23-34.

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepano, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*, 2(2): e0000198. doi.org/10.1371/journal.pdig.0000198

Lim, W.M., Gunasekara, A., Pallant, J.L., Pallant, J.I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarok or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, 21, 100790. DOI: 10.1016/j.ijme.2023.100790

Malinka, K., Peresini, M., Firc, A., Hujnak, O., & Janus, F. (n.d.). On the Educational Impact of ChatGPT: Is Artificial Intelligence Ready to Obtain a University Degree? https://arxiv.org/abs/2303.11146

O'Connor, S., & ChatGPT. (2023). Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Education in Practice*, 66.

OpenAI. (2022). ChatGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/

Pavlik, J.V. (2023). Collaborating with ChatGPT: Considering the implications of Generative Artificial Intelligence for journalism and media education. *Journalism & Mass Communication Educator*, 78(1), 84-93, DOI: 10.1177/10776958221149577

Rospigliosi, P. (2023). Artificial intelligence in teaching and learning: what questions should we ask of ChatGPT?, *Interactive Learning Environments*, 31:1, 1-3, DOI: 10.1080/10494820.2023.2180191

Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays — should professors worry? *Nature*. https://doi.org/10.1038/d41586-022-04397-7.

The Guardian (2023). New York City school bans AI chatbot ChatGPT. https://www.theguardian.com/us-news/2023/jan/06/new-york-city-schools-ban-ai-chatbot-chatgpt

Zhai, X. (2022). ChatGPT user experience: Implications for education. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4312418