International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Describing the Artificial Intelligence's Accuracy in Answering Accounting Problems

Anthony S. Badiang[1], Elite Kaye F. Benasahan[2], Sharam T. Simpal[3], Artchelene D. Pepania[4], Eligen H. Sumicad Jr.[5], Mylene P. Alfanta[6]

[1]Undergraduate Student, Saint Columban College, Pagadian City, Philippines - 7016, anthonybadiang@gmail.com
[2]Undergraduate Student, Saint Columban College, Pagadian City, Philippines - 7016, elitekayefb@gmail.com
[3]Undergraduate Student, Saint Columban College, Pagadian City, Philippines - 7016, stsimpal01@gmail.com
[4]Faculty Member, Saint Columban College, Pagadian City, Philippines – 7016, artchelenepepania@gmail.com
[5]Faculty Member, Saint Columban College, Pagadian City, Philippines – 7016, eligensumicadjr@gmail.com
[6]Faculty Member, Saint Columban College, Pagadian City, Philippines – 7016, mypiegaalfanta@gmail.com

Abstract— *The rise of Artificial Intelligence (AI) significantly impacts the education system by bringing opportunities to expand the frontiers of knowledge. This study is a quantitative descriptive research design that assessed the accuracy level of the five artificial intelligence chatbots in answering accounting problem-based assessments. This study found that Bing Chat performed exceptionally accurate of most accounting courses. Additionally, Bard and ChatGPT just met the expected accuracy, while ChatSonic and Perplexity AI barely met the expected accuracy. Bing Chat, Bard, and ChatGPT render reliable information for the accounting courses, but ChatSonic and Perplexity AI still need to verify and authenticate their credibility and integrity in the accounting courses. Thus, this study recommends to the teachers and the institutions to use Bing Chat, Bard and ChatGPT to help students better assist in their understanding the topics and solving problem-based accounting assessments.*

Keywords— *Accounting course: accounting problems: accuracy: artificial intelligence: Bard: Bing Chat: ChatGPT.*

I. INTRODUCTION

Problem-based learning is mainly known in accounting assessments. Problem-based assessment examines one's broader set of application skills and puts learning on a higher plane (Dockter, 2012). Hugely from COVID-19 pandemic, students have primarily relied on digital technology as online exams became the norm for academic evaluation (Gorgani & Shabani, 2021). Artificial intelligence, one prevalent innovation in society now, has become the students' companion because it can answer questions and solve complex problems like humans (Verma, 2018).

In the literature of Bendal et al. (2020), artificial intelligence is among the most concerning innovations as it begins to impact the lives of the people, the economy, and the world in various forms. The Philippines, country whose digital and high technology systems were less developed, has now dived into the era of AI. According to Ibrahim (2022), the Philippines is among the first 50 countries in the world that have launched the National AI Strategy. The rise of AI significantly impacts the education system of the Philippines by bringing opportunities to expand the frontiers of knowledge. According to Adeva (2023), AI has the potential to unlock academic progress at institutions on a scale never seen before. University of the Philippines (UP) is the first educational institution in the country to set guidelines on AI use in the academe.

In particular, some chatbots may not be accounting-specific but may contain related resources. According to Wood et al. (2023), the chatbot's accurate answers to accounting assessment questions determine whether the chatbot has successfully carried out a positive or negative task. Despite the risk of AI, the country's premier university did not mention banning the technology and issued general guidelines for "responsible" use instead (CNN Philippines, 2023).

The function of AI, specifically chatbots in education, should be ethically and carefully integrated to ensure that quality in learning and assessment is still there. Both humans and AI can learn, but the ability to create knowledge instead of just generating it is what sets them apart. AI chatbots have weaknesses which are their inaccuracy and fabrication of certain information. One research on artificial intelligence chatbots, which demonstrated the correctness of answering the accounting assessment questions, has been used to support this concept.

II. METHODS

The study was conducted in an environment known to be one of the leading institutions in Pagadian City, Philippines. It uses a descriptive - quantitative research design.

In this research study, five (5) Artificial Intelligence (AI) tools were utilized: Bard, Bing Chat, ChatGPT, Chatsonic, and Perplexity AI, which may be software applications or websites. There were various AI tools exists, but mostly are tailored to a specific field of work. The researchers carefully selected AI technologies to be integrated, particularly for this study, that can answer problem-based accounting assessments. The researchers initially tested the chosen AI tools.

The ten (10) Accounting Courses used in this study were problem-based accounting courses with five items per course, the generated answers per item of each course are rated in three criteria, the principles, appropriateness of procedures, and solution and answer. This criterion is a researchers-modified rubric from Sugrue (1995) theory-based framework for assessing domain-specific problem-solving ability. This analytical rubric is presented in two dimensions, with achievement levels as columns and assessment criteria as rows. It evaluates research subjects' achievements using a single rubric based on multiple criteria (Babin & Harrison, 1999). The modified rubric uses an analytical rubric with four scales: 1 - does not meet expectations; 2 - barely meets the expectation approaches; 3 - meets the expectations approaches; and 4 - exceeds the expectations.

The instructors of the said courses provided the five items per course with the assessment of the difficulty level per item. The accuracy of the answers generated by AIs was evaluated by the experts where three (3) experts evaluated each accounting course.

III. Results and Discussions

A. Bard

TABLE 1: Accuracy Level of Bard

Courses	Mean	SD	Interpretation
Accounting for Business Combination	2.33	0.3685	Barely meets the expectation approaches
Auditing and Assurance Concepts and Applications I	2.47	0.3083	Barely meets the expectation approaches
Auditing and Assurance Concepts and Applications II	3.11	0.2722	Meets the expectation approaches
Cost Accounting and Control I	2.58	0.3881	Meets the expectation approaches
Financial Management	2.80	0.3083	Meets the expectation approaches
Intermediate Accounting I	2.71	0.4818	Meets the expectation approaches
Intermediate Accounting II	3.20	0.7175	Meets the expectation approaches
Intermediate Accounting III	2.60	0.5247	Meets the expectation approaches
Strategic Cost Management	2.76	0.7470	Meets the expectation approaches
Valuation, Concepts and Methods	3.62	0.3388	Exceeds the expectation
Grand Mean	2.82	2.82	Meets the expectation approaches

The experts agreed that Bard shows the highest level of accuracy in the Valuation, Concepts, and Methods course. It means that Bard identifies significant principles and standards theories in accounting accurately, applies appropriate concepts and procedures, and demonstrates solutions accurately and relevant answers. Bard is much better in handling questions that have been already published and answered online, than original questions that have not yet been published. Plevris et al. (2023) stated in their findings that the published questions were, in fact, harder than the original questions, and Bard has direct access to the internet, which is Google's search engine.

B. Bing Chat

All items regarding Accounting for Business Combinations course have the highest rating amongst the ten accounting courses manifesting the overall mean response of 3.58 with SD of 0.6106 that is interpreted as exceeds the expectation approaches. Following the highest rating in the accuracy of Bing Chat are Intermediate Accounting I with overall mean of 3.51 and SD of 0.4554, Valuation, Concepts and Methods which got an overall mean of 3.51 and SD 0.462, and Intermediate Accounting II manifesting the overall mean response of 3.40 with SD of 0.5698, which are all interpreted as exceeds the expectation approaches. Bing Chat performed poorly in Cost Accounting and Control I, manifesting the overall mean response of 2.29 and SD of 0.3201 which is interpreted as barely meets the expectation approaches. The grand mean of Bing Chat is 3.16, which is construed as meets the expectation approaches. Due to its use of the more sophisticated GPT-4 technology, which enables it to comprehend and reply to complicated questions and provide more accurate and detailed responses, Bing Chat displayed higher performance across a variety of accounting disciplines. (Dao & Le, 2023). Bing Chat's effectiveness in delivering precise responses and informative explanations in various accounting courses renders it a valuable resource for accounting students seeking support in those array of accounting courses.

TABLE 2: Accuracy Level of Bing Chat

Courses	Mean	SD	Interpretation
Accounting for Business Combination	3.58	0.6106	Exceeds the expectation
Auditing and Assurance Concepts and Applications I	3.20	0.5741	Meets the expectation approaches
Auditing and Assurance Concepts and Applications II	3.13	0.9040	Meets the expectation approaches
Cost Accounting and Control I	2.29	0.3201	Barely meets the expectation approaches
Financial Management	2.87	0.7754	Meets the expectation approaches
Intermediate Accounting I	3.51	0.4554	Exceeds the expectation
Intermediate Accounting II	3.40	0.5698	Exceeds the expectation
Intermediate Accounting III	3.24	0.7001	Meets the expectation approaches
Strategic Cost Management	2.89	0.5827	Meets the expectation approaches
Valuation, Concepts and Methods	3.51	0.4621	Exceeds the expectation
Grand Mean	3.16		Meets the expectation approaches

C. ChatGPT

TABLE 3: Accuracy Level of ChatGPT

Courses	Mean	SD	Interpretation
Accounting for Business Combination	2.82	0.6266	Meets the expectation approaches
Auditing and Assurance Concepts and Applications I	2.58	0.9539	Meets the expectation approaches
Auditing and Assurance Concepts and Applications II	2.82	0.6555	Meets the expectation approaches
Cost Accounting and Control I	2.09	0.2137	Barely meets the expectation approaches
Financial Management	2.82	0.6507	Meets the expectation approaches
Intermediate Accounting I	2.84	0.3201	Meets the expectation approaches
Intermediate Accounting II	3.38	0.6315	Exceeds the expectation
Intermediate Accounting III	2.71	0.8035	Meets the expectation approaches
Strategic Cost Management	2.73	0.7601	Meets the expectation approaches
Valuation, Concepts and Methods	3.09	0.6592	Meets the expectation approaches
Grand Mean	2.79		Meets the expectation approaches

According to the information given, ChatGPT is generally only performing up to expectations, which would be considered a success for the accounting courses and place a student academically in good standing. ChatGPT's marks would only be acceptable for a student to pass the courses if its performance was consistent across the accounting curriculum. As seen by its virtually exceptional performance, ChatGPT generally scored at or meeting the expectation approaches, which interprets that the results identify or apply some principles and concepts of accounting, demonstrating answers with some accuracy and relevance. ChatGPT received the highest rating in Intermediate Accounting II amongst the ten problem-based accounting courses, manifesting a mean response of 3.38 and SD of 0.6315, which is interpreted as exceeding the expectation approaches, while it barely meets the expectations on Cost Accounting and Control I, which has the lowest mean response of 2.09 with SD of 0.2137. The findings of Wood et al. (2023) revealed that even when ChatGPT's answers were incorrect, it frequently gave thorough justifications. This raises the crucial question of how students might be impacted by these authoritative but inaccurate responses. When answering accounting queries for which accounting standards have remained constant over time and require less judgment, ChatGPT performed well. Long, complex written questions were another challenge for ChatGPT.

D. Chatsonic

TABLE 4: Accuracy Level of Chatsonic

Courses	Mean	SD	Interpretation
Accounting for Business Combination	2.31	0.4332	Barely meets the expectation approaches
Auditing and Assurance Concepts and Applications I	2.44	0.4969	Barely meets the expectation approaches
Auditing and Assurance Concepts and Applications II	2.60	0.9895	Meets the expectation approaches
Cost Accounting and Control I	2.24	0.2137	Barely meets the expectation approaches
Financial Management	2.98	0.6592	Meets the expectation approaches
Intermediate Accounting I	2.27	0.2897	Barely meets the expectation approaches
Intermediate Accounting II	3.31	0.9441	Exceeds the expectation
Intermediate Accounting III	2.49	0.9895	Barely meets the expectation approaches
Strategic Cost Management	2.11	0.2485	Barely meets the expectation approaches
Valuation, Concepts and Methods	3.40	0.8300	Exceeds the expectation
Grand Mean	2.62		Meets the expectation approaches

All items regarding the Valuation, Concepts, and Methods course have the highest rating amongst the ten accounting courses, manifesting the overall mean response of 3.40 and SD of 0.8300, interpreted as exceeding the expectation approaches. Following the highest rating with the interpretation as exceeds the expectation approaches is the Intermediate Accounting II with a mean of 3.31 and a standard deviation of 0.9441. Although ChatSonic displayed a grand mean of 2.62, which is construed as meeting the expectation approach, it is essential to note that it barely meets the expectation of accuracy in most accounting courses. Chaka (2023) implied in his study that without crediting the sources, ChatSonic displayed a propensity to copy responses from online content, leading to rambling responses in some places. As a result, ChatSonic is yet to be a trustworthy and reliable source of information for accounting courses.

E. Perplexity AI

TABLE 5: Accuracy Level of Perplexity AI

Courses	Mean	SD	Interpretation
Accounting for Business Combination	2.67	0.7158	Meets the expectation approaches
Auditing and Assurance Concepts and Applications I	2.40	0.4202	Barely meets the expectation approaches
Auditing and Assurance Concepts and Applications II	2.09	0.6256	Barely meets the expectation approaches
Cost Accounting and Control I	2.07	0.2018	Barely meets the expectation approaches
Financial Management	2.84	0.6507	Meets the expectation approaches
Intermediate Accounting I	2.02	0.8066	Barely meets the expectation approaches
Intermediate Accounting II	3.18	0.8484	Meets the expectation approaches
Intermediate Accounting III	2.38	0.6363	Barely meets the expectation approaches
Strategic Cost Management	1.89	0.7370	Barely meets the expectation approaches
Valuation, Concepts and Methods	3.16	0.7478	Meets the expectation approaches
Grand Mean	2.47		Barely meets the expectation approaches

The grand mean of Perplexity AI has a result of 2.47 and is interpreted as barely meets the expectation. As indicated in the study of Fostikov (2023), the test version of Perplexity AI was released and is still being refined. As a result, this chatbot still needs to be checked over and fixed because it sometimes provides imprecise responses.

Overall, the level of accuracy of Bard, Bing Chat, ChatGPT, ChatSonic, and Perplexity AI is highest in the Valuation, Concepts, and Methods course, while it is lowest in the Cost Accounting and Control I course. Bing Chat performed exceptionally among the five AI chatbots as it has the highest grand mean of 3.16, meeting the expectation of accuracy of most accounting courses. Bard and ChatGPT also met the expectation of accuracy with a grand mean of 2.82 and 2.79, respectively. The researchers wanted also to imply that the ratings obtain by AI may include the factors on the standards it follows in generating the answers. Philippine Accounting Standards (PAS), Philippine Financial Reporting Standards (PFRS), interpretations, pronouncements and other relevant Philippine laws are the standards overriding in all facets of accounting. Although PAS and PFRS corresponds to the adopted International Accounting Standards and International Financial Accounting Standards, respectively, some pertinent framework in the Philippines may differ from the foreign countries. It is likely to have a difference in rules of applications and implementation in the Philippines and in foreign countries. These AIs does not program to follow certain standards in application. ChatSonic and Perplexity AI displayed the grand mean of 2.62 and 2.47, respectively, implying the need to verify and authenticate the credibility and integrity of the information provided by these AIs.

IV. CONCLUSIONS

The Bing Chat, Bard, and ChatGPT render reliable information in solving accounting problem courses discussed in this paper. While currently inconsistent at best in all courses, Bing Chat, ChatGPT, and Bard's performance in accounting courses suggests both great promise and peril. The researchers found that these language models will be crucial for future accounting practice and beneficial for students using them in accountancy programs.

REFERENCES

[1] A. Bendal, S. A. Planas, and R. G. Atento. “Impact of artificial intelligence as a disruptive technology on accountancy,” LPU-Laguna Journal of Business and Accountancy, vol. 3, issue 3, 2020.

[2] A. Fostikov. “First impressions on using AI powered chatbots, tools and search engines: ChatGPT, Perplexity and other – possibilities and usage problems,” Review of the NCD, pp. 12–21, 2023.

[3] B. Sugrue. “A theory-based framework for assessing domain-specific problem-solving ability,” Educational Measurement: Issues and Practice, vol. 14, issue 3, pp. 29–36, 1995.

[4] C. Chaka. “Generative AI chatbots - ChatGPT versus YouChat versus Chatsonic: Use cases of selected areas of applied English language studies,” International Journal of Learning, Teaching and Educational Research, vol. 22, issue 6, pp. 1-19, 2023.

[5] CNN Philippines. “UP crafts guidelines for 'responsible' AI use,” 2023.

[6] D. A. Wood, M. P. Achhpilia, M. T. Adams, S. Aghazadeh, K. Akinyele, M. Akpan, K. D. Allee, A. M. Allen, E. D. Almer, D. Ames, V. Arity, D. Barr-Pulliam, K. A. Basoglu, A. Belnap, J. W. Bentley, T. Berg, N. R. Berglund, E. Berry, A. Bhandari, M. N. H. Bhuyan, P. Black, ... E. Zoet. “The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions?” Issues in Accounting Education, vol. 38, issue 4, pp. 81-108, 2023.

[7] D. L. Dockter. “Problem-based learning in accounting,” American Journal of Business Education, vol. 5, issue 5, 2012.

[8] E. Babin and K. Harrison, Contemporary Composition Studies, 1st ed. Westport, 1999.

[9] H. H. Gorgani and S. Shabani. “Online exams and the COVID-19 pandemic: A hybrid modified FMEA, QFD, and K-means approach to enhance fairness,” SN Applied Sciences, vol. 3, issue 10, 2021.

[10] M. Ibrahim. “Artificial intelligence in the Philippines,” Manila Bulletin, 2022.

[11] M. Verma. “Artificial intelligence and its scope in different areas with special reference to the field of education,” International Journal of Advanced Educational Research, vol. 3, issue 1, pp. 5-10, 2018.

[12] P. K. Adeva, “Will AI chatbots take over education?” Philippine Collegian, 2023.

[13] V. Plevris, G. Papazafeiropoulos, and A. J. Rios. “Chatbots put to the test in math and logic problems: A comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard,” AI, vol. 4, issue 4, pp. 949-969, 2023.

[14] X. Q. Dao and N. B. Le, “ChatGPT is good but Bing Chat is better for Vietnamese Students,” 2023.