R E V I S T A
Globalización, Competitividad y Gobernabilidad

J O U R N A L
Globalization, Competitiveness & Governability

R E V I S T A
Globalização, Competitividade e Governabilidade

GCG
The Journal *of* Globalization,
Competitiveness, and Governability

Application of Chatgpt 4.0 for Solving Accounting Problems

AREA: 2
TYPE: Case

Aplicação do Chatgpt 4.0 para Resolução de Problemas Contábeis
Aplicación de Chatgpt 4.0 para Solucionar Problemas Contables

AUTHORS

Marcelo Machado de Freitas
Universidade Federal de Santa Catarina (UFSC), Brasil
mmf.marcelofreitas@gmail.com

Jonatas Dutra Sallaberry[1]
Universidade do Contestado, Brasil
jonatas.sallaberry@hotmail.com

Thiago Bruno de Jesus Silva
Universidade Federal do Recôncavo da Bahia (UFRB), Brasil
thiagobruno.silva@yahoo.com.br

Fabrícia Silva da Rosa
Universidade Federal de Santa Catarina (UFSC), Brasil
fabriciasrosa@hotmail.com

In the context of the accounting professional's problems and tasks that can be optimized with Artificial Intelligence applications, the research aimed to identify and evaluate the performance of the ChatGPT tool, model GPT-4.0, in resolving questions from the Accounting Sufficiency Exam, as a proxy for problems in accounting activity. For this verification, we applied questions from four editions of the Exam in ChatGPT 4.0. The results indicate approval in the four editions in which only 20 to 23% of candidates were approved, while the detailed performance showed 71% of ChatGPT's correct answers compared to 44% of the candidates.

No contexto dos problemas e tarefas do profissional contábil que podem ser otimizados com aplicações de Inteligência Artificial, a pesquisa teve por objetivo identificar e avaliar o desempenho da ferramenta ChatGPT, modelo GPT-4, na resolução de questões do Exame de Suficiência em Contabilidade, como proxy de problemas na atividade contábil. Para essa verificação aplicamos as questões de quatro edições do Exame no ChatGPT 4. Os resultados indicam aprovação nas quatro edições em que apenas 20 a 23% dos candidatos foram aprovados, enquanto o desempenho detalhado apontou 71% de acertos do ChatGPT frente a 44% dos candidatos.

En el contexto de los problemas y tareas del profesional contable que pueden optimizarse con aplicaciones de Inteligencia Artificial, la investigación tuvo como objetivo identificar y evaluar el desempeño de la herramienta ChatGPT, modelo GPT-4.0, en la resolución de preguntas del Examen de Suficiencia Contable, como un indicador de problemas en la actividad contable. Para esta verificación, aplicamos preguntas de cuatro ediciones del Examen en ChatGPT 4.0. Los resultados indican aprobación en las cuatro ediciones en las que solo entre el 20 y el 23% de los candidatos fueron aprobados, mientras que el desempeño detallado mostró el 71% de las respuestas correctas de ChatGPT en comparación con 44% de cuestiones correctas de los candidatos.

DOI
10.58416/GCG.2024.V18.N2.03

Received
19.09.2023

Accepted
02.11.2023

---

[1] Corresponding Author. Avenida Nereu Ramos, 1071, Jardim do Moinho, 89300000 - Mafra, SC - Brasil

1. Introduction

The ever-expanding use of Information and Communication Technologies (ICTs) is one of the most relevant contemporary developments for companies, employees, and society (Lara, 2023). Particularly noteworthy in the field of Artificial Intelligence (AI) is the emergence and refinement of Natural Language Processing (NLP) models, such as ChatGPT 4.0, capable of generating text in a manner similar to humans (Brown et al., 2020).

AI can be defined as the ability of a computer to perform processes that simulate human intelligence, reasoning to understand data and their relationships, as well as self-correction mechanisms to make decisions (Mujiono, 2021). For this reason, AI has garnered interest across various fields of knowledge, including accounting (AICPA, 2019). AI models have the potential to transform the accounting profession as these technologies are integrated into traditional accounting processes, assisting in financial report preparation, audit processes (Chui et al., 2016), automating manual tasks, and enhancing decision support (Chui et al., 2016, Vasarhelyi et al., 2015).

Recent research has analyzed the performance of AI systems in professional examinations (Bordt & Von Luxburg, 2023; Choi et al., 2023; Gilson et al., 2023; Subramani et al., 2023), demonstrating the capability of these models to solve problems across diverse areas. Specifically in accounting, the ChatGPT model (GPT 3.5) failed to answer questions in the Certified Public Accountant (CPA) exam, which qualifies individuals for the accounting profession in the United States, with a success rate between 35% to 48% (Accounting Today, 2023). However, the most recent model of ChatGPT (GPT 4) outperformed its previous versions, showing substantially improved performance in certain professions' proficiency exams (Bommarito, 2023; Katz et al., 2023; Martínez et al., 2023).

Given this context, the research question arises: **what is the performance of ChatGPT, GPT-4 model, in the Accounting Proficiency Exam in Brazil?** By addressing this question, the study aims to identify and evaluate the performance of the ChatGPT tool, GPT-4 model, in solving questions from the Accounting Proficiency Exam, as a proxy for problems in accounting activities.

While theoretical or applied questions in accounting content may not fully reflect the capacity for advanced task execution by accounting professionals, the decision was made to use the questions from the Proficiency Exam as a proxy for this variable. This choice is justified by its relevance, as passing the exam is a legal requirement for practicing accounting in Brazil.

KEYWORDS
Artificial Intelligence, ChatGPT, Tasks.

PALAVRAS-CHAVE
Inteligência Artificial, ChatGPT, Tarefas.

PALABRAS CLAVE
Inteligencia Artificial, ChatGPT, Tareas.

JEL CODES
M15, M41, M54

2. Theoretical Framework

The accounting profession has undergone numerous transformations in recent decades, driven by business globalization, diversified activities, generational characteristics, but the main determinant emerges as the technology applied to individuals and businesses (Lara, 2023). Among the most advanced technologies is artificial intelligence through Natural Language Processing (NLP) models. These models are capable of performing a diverse range of tasks, whether simple or complex, with reasonable accuracy (Else, 2023; Liu et al., 2021).

NLP models have undergone significant improvements in recent years, primarily due to advancements in deep learning and neural networks, alongside enhancements in the processing performance of these models (Liu et al., 2023). These advancements led to the development of more sophisticated models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT, which have outperformed previous models in various NLP tasks, including sentiment analysis, responding to different questions, and tasks associated with text classification, among others (Kasneci et al., 2023; Kenton & Toutanova, 2019; Radford et al., 2019).

Furthermore, recent empirical evidence supports the opportunity to enhance the performance of NLP models through external queries, notepad blocks, Chain of Thought (CoT) prompts, or one of the many other techniques that frequently emerge (Bommarito, 2023; Katz et al., 2023; Martínez et al., 2023). Among the various tools offering NLP models for diverse uses is ChatGPT, which has rapidly gained popularity but still raises ethical questions (Alves, 2023).

2.1. What is ChatGPT and how NLP models work

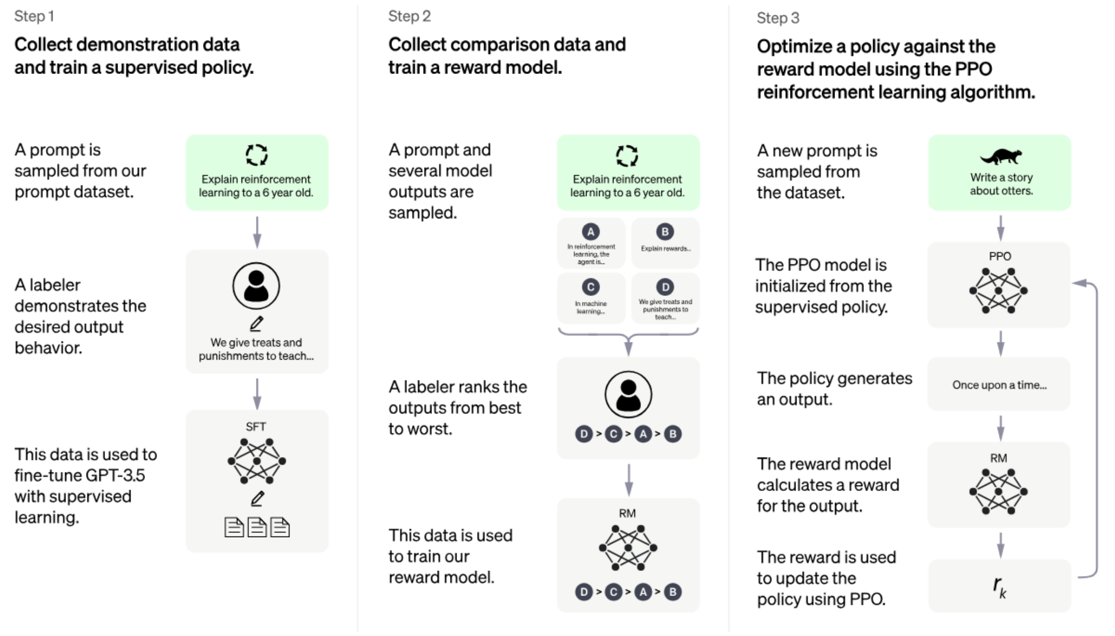
ChatGPT, or Chat-based Generative Pre-trained Transformer, is a language model developed by OpenAI that uses a large amount of textual data to generate human-like conversational responses (Radford et al., 2018).

The primary innovation behind GPT, GPT-4, and other models like BERT lies in their pre-training on extensive textual data, which helps the model learn the structure and semantics of language (Kasneci et al., 2023; Kenton & Toutanova, 2019), allowing it to generate responses that resemble human-generated text (Radford et al., 2019). The model architecture consists of a series of self-attention layers that assist in analyzing and understanding the input text or prompt in the case of ChatGPT. These layers enable the model to identify important relationships between words, even if they are distant, which is useful for capturing the overall meaning of a text or problem statement and better understanding the structure and grammar of language (Vaswani et al., 2017).

The latest versions of the AI application ChatGPT include the GPT-3 and GPT-4 models (the most current at the time of research). GPT-3 is an autoregressive language model with 175 billion parameters, ten times more than any previous language model (Brown et al., 2020). Meanwhile, the GPT-4 model is a more advanced model that has demonstrated accurate performance in various NLP tasks (Brown et al., 2020; Else, 2023).

ChatGPT was trained using Reinforcement Learning and Human Feedback (RLHF) models. Initially, a model was trained using supervised fine-tuning, where AI trainers provided conversations playing both roles, as a user and an AI assistant. Trainers had access to suggestions written by the model to aid in formulating responses. Subsequently, these dialogues were mixed with the InstructGPT dataset, which was then transformed into a dialogue format, resulting in the current structure of ChatGPT (OpenAI, 2023).

Figure 1. - Explanatory steps of ChatGPT's NLP functioning


Step 1
Collect demonstration data and train a supervised policy.
A prompt is sampled from our prompt dataset.
Explain reinforcement learning to a 6 year old.
A labeler demonstrates the desired output behavior.
We give treats and punishments to teach...
This data is used to fine-tune GPT-3.5 with supervised learning.
SFT
Step 2
Collect comparison data and train a reward model.
A prompt and several model outputs are sampled.
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
A labeler ranks the outputs from best to worst.
D > C > A > B
This data is used to train our reward model.
RM
D > C > A > B
Step 3
Optimize a policy against the reward model using the PPO reinforcement learning algorithm.
A new prompt is sampled from the dataset.
Write a story about otters.
The PPO model is initialized from the supervised policy.
PPO
The policy generates an output.
Once upon a time...
The reward model calculates a reward for the output.
RM
The reward is used to update the policy using PPO.
r_k

Source: OpenAI (2023)

Some uses of ChatGPT have already been analyzed in the academic world. Else (2023), for instance, demonstrated that ChatGPT was able to convincingly write abstracts of scientific articles. According to the author's results, researchers were unable to adequately distinguish between abstracts written by other researchers and those generated by artificial intelligence. These findings demonstrate how NLP models can be integrated into the daily routines of various professions, including accounting, assisting or even replacing several professional tasks.

However, these models are not free from issues. According to OpenAI (2023), models like ChatGPT can generate false, toxic, or reflect harmful sentiments. This happens because GPT-3 and similar models were trained to predict the next word in a large dataset of internet text, rather than securely performing the linguistic task that the user usually requires.

ChatGPT-4, as an NLP model, benefits from recent advancements in the field and enhances them to provide a wide range of applications, including language translation, summarization, and natural language understanding (Brown et al., 2020). Its ability to generate coherent, context-aware, and relevant responses makes it an important resource in various fields, including accounting.

2.2. Potentials of AI and NLP to influence accounting

AI and NLP models have the potential to revolutionize various aspects of accounting, from automating repetitive tasks to enhancing decision-making processes (Davenport & Kirby, 2016). According to Chui et al. (2016), a considerable portion of activities performed by finance professionals can be automated in the short term, especially concerning data collection and processing.

The utilities of AI include problem-solving, natural language processing, speech recognition, image processing, automated programming, and robotics (Amaral et al., 2023). These technologies are capable of storing and processing data much more effectively and efficiently than a human, thereby enabling the substitution of workers by machines in some processes and the reallocation of employees to more essential or sensitive business activities (Mancebo & Mucci, 2023).

NLP models, in particular, can extract valuable information from unstructured data such as financial statements, annual reports, and auditor's notes (Loughran & McDonald, 2016). By incorporating historical data and real-time financial information, these Machine Learning algorithms can identify trends and patterns that may not be easily perceptible to human analysts, leading to more accurate predictions and well-founded strategic decisions (Makridakis et al., 2018).

These technologies can assist in identifying opportunities in the accounting field, tax savings, and optimizing tax declarations, making accounting more efficient (Kogan et al., 2017). Accounting education, as an accounting environment in constant exchange with the profession, is a field that can also benefit from AI. AI or NLP-based tutors can provide personalized guidance to accounting students, allowing them to learn at their own pace and address specific areas where they might have weaknesses (Baidoo-Anu et al., 2023).

NLP models like GPT-4 can be used to facilitate communication between students and instructors, providing instant feedback and clarifications on complex concepts, as demonstrated in other fields of knowledge such as mathematics (Pardos & Bhandari, 2023) and engineering (Qadir, 2022). For instance, ChatGPT has already been used to analyze sufficiency exams for lawyers (Bommarito, 2023; Katz et al., 2023; Martínez et al., 2023).

While GPT-4 has demonstrated the ability to learn the structure and semantics of language (Kasneci et al., 2023; Kenton & Toutanova, 2019), enabling it to generate responses similar to human-generated texts, the model may still make mistakes such as hallucinating sources, misinterpreting facts, or failing to follow ethical requirements. Given the rapid evolution in the use of these models in daily life, the need for scientific studies that monitor the tool's expansion and its utility across various professions becomes crucial (Katz et al., 2023).

Consequently, considering the Accounting Proficiency Exam as a proxy for the accounting knowledge that enables graduates from the undergraduate course to practice the profession has proven useful. This can be an initial demonstration of AI capabilities in addressing theoretical and practical problems that can be answered in descriptive terms within the accounting world.

3. Methodological Procedures

The comparison of AI with the Accounting Proficiency Exam is intriguing when measuring the knowledge acquired by future accounting professionals because it has restricted the professional qualification of a significant number of graduates from Accounting courses, owing to the candidates' low performance in this Exam. For instance, out of the 35,984 candidates registered and present for the 2022/2 Exam, only 7,595 achieved approval, which is 21.11% (CFC, 2023), as shown in **Table 1**.

Table 1 - Statistics by Content

Content	*Q*	*2022 Sem 2*	*2022 Sem 1*	*2021 Sem 2*	*2021 Sem 1*	*2020 Sem 2*	*2020 Sem 1*	*2019 Sem 2*	*2019 Sem 1*	*2018 Sem 2*	*2018 Sem 1*	*Average*
Audit	2	37,69	84,04	39,68	16,37	19,38	52,29	52,99	42,16	44,50	57,97	44,71
Accounting for Public Sector	3	27,11	45,63	44,25	36,64	44,14	38,94	29,50	53,71	36,20	22,35	37,85
Cost Accounting	4	33,26	29,28	24,00	39,59	46,32	44,23	25,70	52,16	47,62	22,70	36,49
General Accounting	17	42,96	40,91	48,33	37,19	39,41	51,74	40,73	41,19	56,99	49,70	44,92
Management	2	33,24	18,08	26,94	19,31	20,73	46,72	29,90	40,83	41,63	35,80	31,32
Controllership	1	16,76	48,48	54,62	100,00	40,99	100,00	12,74	27,47	32,27	66,41	49,97
Leg & Prof Ethics	4	62,74	59,72	74,98	79,81	69,95	67,61	70,87	80,72	54,86	53,05	67,43
Portuguese Language	2	24,87	22,65	42,39	46,80	40,87	29,72	35,98	38,75	45,34	26,45	35,38
Fin Math & Stats	2	47,44	17,80	26,95	24,27	27,08	32,80	33,28	44,42	26,27	26,99	30,73
Law & Applied Leg.	3	36,98	46,12	46,88	69,27	39,62	42,36	72,44	56,97	51,55	60,33	52,25
Forensic Acctg	2	62,24	31,90	34,52	39,44	76,59	67,94	55,75	29,94	31,14	38,59	46,81
Acctg Principles & BRGAAP	4	38,35	31,80	25,38	30,82	45,74	29,89	60,22	33,68	30,11	14,79	34,08
Acctg Theory	4	52,53	66,12	51,55	39,15	38,81	44,62	48,88	46,79	33,03	67,65	48,91

Source: CFC (2023)

These results also reveal discrepancies among distinct content groups (Miranda et al., 2017), as depicted in **Table 1**, showing the quantity of questions per knowledge area, as well as the percentage performance of examinees from 2018 to 2022 and the performance average in the period.

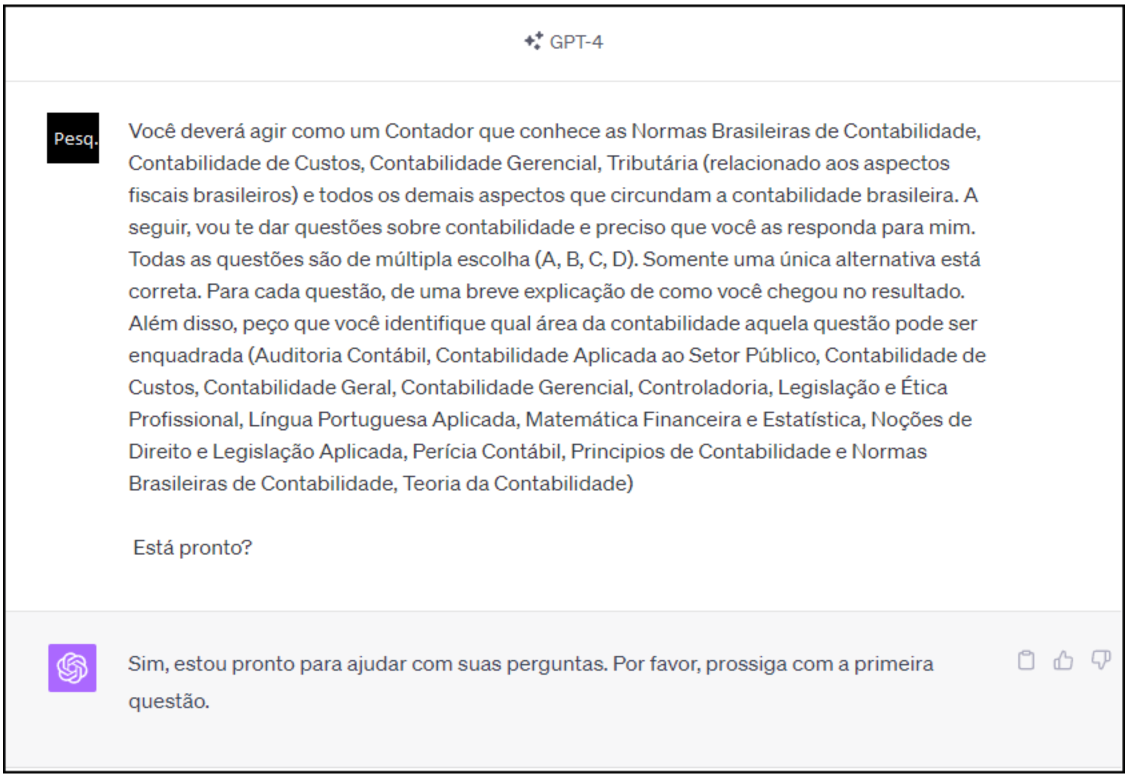
To address the research objective, questions from the proficiency exam from four editions (2021 and 2022) were extracted and tabulated with their respective answer options. Subsequently, these questions were manually inputted into ChatGPT, GPT-4 model (https://chat.openai.com) using one of the researchers' accounts for this study.

The research was operationalized using the Design Science methodology, aiming to construct and evaluate different artifacts (technological or non-technological) within a specific field of knowledge (Hevner et al., 2004). Hevner et al. (2004) explain that the Design Science paradigm seeks to expand the limits of human and organizational capabilities by creating and evaluating innovative artifacts.

The application of questions from the Accounting Proficiency Exam simulating routine accounting problems follows previous studies (Bommarito, 2023; Katz et al., 2023; Martínez et al., 2023), which not only demonstrated the evolution of ChatGPT-4 but also the relevance of the tool due to linguistic and relational advancements, enabling the evaluation of its application in professional exams (Katz et al., 2023).

For interaction with the technology, based on the question models and the research objective, the authors developed and validated with experts the following conditional question to be inserted as a prompt at the beginning of the conversation (**Figure 2**):

Figure 2. - ChatGPT Prompt – Initial Interaction[2]



Você deverá agir como um Contador que conhece as Normas Brasileiras de Contabilidade, Contabilidade de Custos, Contabilidade Gerencial, Tributária (relacionado aos aspectos fiscais brasileiros) e todos os demais aspectos que circundam a contabilidade brasileira. A seguir, vou te dar questões sobre contabilidade e preciso que você as responda para mim. Todas as questões são de múltipla escolha (A, B, C, D). Somente uma única alternativa está correta. Para cada questão, de uma breve explicação de como você chegou no resultado. Além disso, peço que você identifique qual área da contabilidade aquela questão pode ser enquadrada (Auditoria Contábil, Contabilidade Aplicada ao Setor Público, Contabilidade de Custos, Contabilidade Geral, Contabilidade Gerencial, Controladoria, Legislação e Ética Profissional, Língua Portuguesa Aplicada, Matemática Financeira e Estatística, Noções de Direito e Legislação Aplicada, Perícia Contábil, Principios de Contabilidade e Normas Brasileiras de Contabilidade, Teoria da Contabilidade)

Está pronto?

Sim, estou pronto para ajudar com suas perguntas. Por favor, prossiga com a primeira questão.

Source: authors (2023) and ChatGPT 4 (2023)

Subsequently, each question was inserted into the chat, one by one, without any new prompts during execution. In questions that presented tables, data was extracted and presented sequentially by lines. Procedures were conducted to validate if ChatGPT could comprehend the information presented in this format, which was confirmed.

After the resolution of each question, the outputs were extracted into spreadsheets, where each response presented by the chat was inserted and compared with the official answer shared by CFC. In the case of questions invalidated by the CFC and the examining board, the ChatGPT response was considered correct. To fulfill the research scope, the authors categorized the questions according to the contents and their quantities indicated by the Accounting Proficiency Exam organization committee, enabling analysis by content.

The ChatGPT responses were analyzed quantitatively (to demonstrate the number of correct answers per test and by area) and qualitatively (to understand how the responses were structured and the main errors of the model).

4. Results

4.1. General Results

The application of ChatGPT for solving questions from the Accounting Proficiency Exam across four editions demonstrated sufficient resolution capability for professional accounting qualification. While legislation requires a minimum score of 50% in the exam questions, the ChatGPT, GPT-4 model, showed an average score of 71%, fluctuating between 64 to 78%, as shown in **Table 2**.

The analysis of categorized question data by content also allows for a more detailed analysis of the model's efficiency by theme in each exam edition and the overall performance over the consolidated 200 questions.

Table 2 - ChatGPT Results

Content	*Total*	*Correct 21-1*	*Percent 21-1*	*Correct 21-2*	*Percent 21-2*	*Correct 22-1*	*Percent 22-1*	*Correct 22-2*	*Percent 22-2*	*Correct General*	*Percent General*
General Accounting	17	11	65%	13	76%	14	82%	10	59%	48	71%
Cost Accounting	4	4	100%	3	75%	3	75%	2	50%	12	75%
Accounting for Public Sector	3	2	67%	2	67%	3	100%	2	67%	9	75%
Management	2	1	50%	2	100%	2	100%	1	50%	6	75%
Controllership	1	1	100%	0	0%	1	100%	0	0%	2	50%
Law & Applied Leg.	3	2	67%	2	67%	3	100%	3	100%	10	83%
Fin Math & Stats	2	2	100%	2	100%	0	0%	2	100%	6	75%
Acctg Theory	4	3	75%	2	50%	3	75%	3	75%	11	69%
Leg & Prof Ethics	4	4	100%	4	100%	3	75%	2	50%	13	81%

Acctg Principles & BRGAAP	4	2	50%	1	25%	1	25%	1	25%	5	31%
Audit	2	0	0%	2	100%	2	100%	2	100%	6	75%
Forensic Acctg	2	1	50%	1	50%	2	100%	2	100%	6	75%
Portuguese Language	2	1	50%	2	100%	2	100%	2	100%	7	88%
Sum	*50*	*34*	*68%*	*36*	*72%*	*39*	*78%*	*32*	*64%*	*141*	*71%*

4.2. Discussions

The highest scores for ChatGPT were in the contents of Applied Portuguese Language, Notions of Law and Applied Legislation, and Professional Ethics and Legislation, respectively with 88, 83, and 81% accuracy. Despite a grouping of 36 verified questions, the technology correctly answered 30 questions, or 83% of the total.

These contents demand theoretical knowledge of a significant number of norms, regulations, resolutions, and laws. Besides cognitive understanding and rationalization, this content type requires the storage of a large amount of text and normative peculiarities, naturally facilitated for digital memory. Even questions related to Portuguese Language require the application of spelling and grammatical rules that demand a high memorization load.

Some plausible explanations for the model's higher accuracy in these types of questions might be that:

i) ChatGPT uses NLP to understand inputs and construct its responses. This means its main job is to process human language. Hence, interpreting texts in Portuguese and analyzing grammatical and cohesion-related questions was something ChatGPT should perform reasonably well, which in fact was identified as being the case;

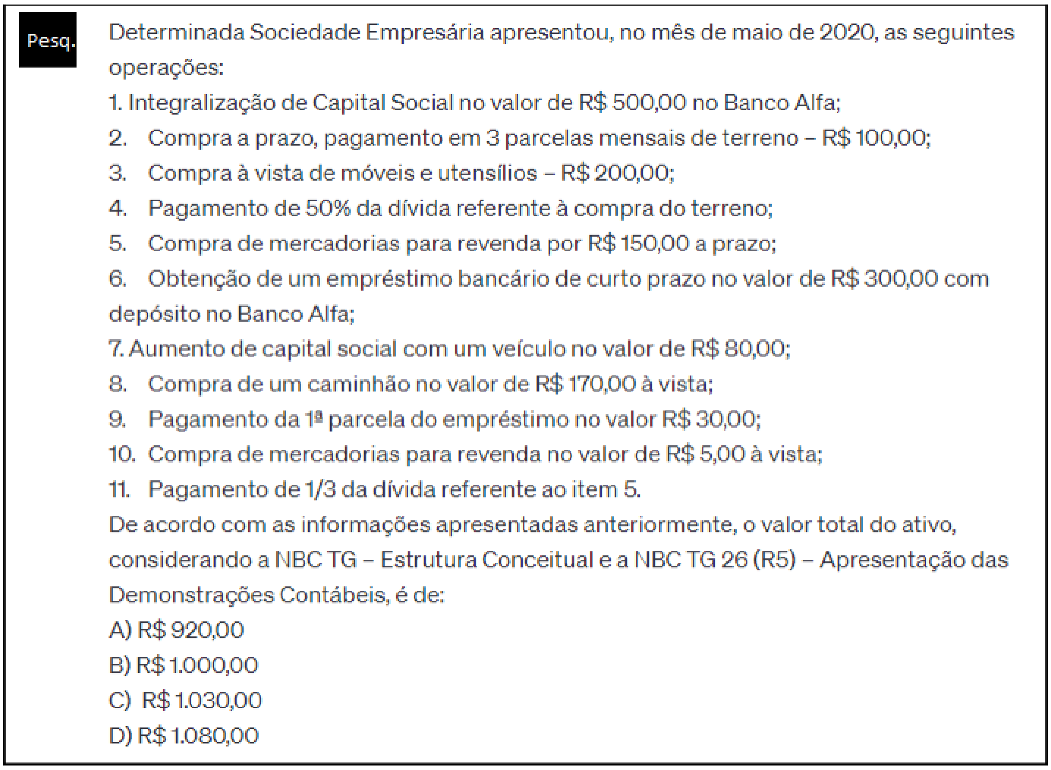
ii) For an NLP model, it's easier to contextualize legislation than to perform mathematical calculations. Several examples have shown that even more recent models, like GPT-4, still face certain difficulties in basic math operations. The reason is that these models were not built for that purpose but, as mentioned earlier, to process language. Therefore, themes involving legislations (with exceptions in highly complex and subjective situations) are possibly more easily answered by these models compared to questions analyzing multiple accounting entries together, for instance.

On the other end of the spectrum, contents with lower accuracy were Accounting Principles and Brazilian Accounting Standards, Controllership, and Accounting Theory, respectively scoring 31, 50, and 69%. The content group also encompasses 36 questions, of which only 18 were correctly answered by ChatGPT, that is, a 50% score, nevertheless, sufficient for passing the exam and professional qualification.

These contents that exhibited lower accuracy for the model represent significantly extensive memorization content, which would be more facilitated for the technology. However, they represent necessary accounting concepts applied to measurement, accounting, and recognition problems, which often require professional judgment. In this empirical research, it was shown that performance in these areas was lower compared to others, but still sufficient for qualification in the exam, and superior to the candidates' average.

These results demonstrate that technology has the capacity to access and present significant knowledge in the accounting field, considering the diverse topics covered and the complexity of the questions, with results surpassing the average of other Accounting graduates. However, considering this technological tool as a mechanism for decision-making knowledge needs to be analyzed cautiously due to interpretation and processing errors, as exemplified in **Figure 3**, from the 2021.01 Proficiency Exam – Question 42, whose official answer is option "B) R$ 1,000.00."

Figure 3 - ChatGPT Prompt – Example of Inquiry[3]



Pesq.

Determinada Sociedade Empresária apresentou, no mês de maio de 2020, as seguintes operações:

1. Integralização de Capital Social no valor de R$ 500,00 no Banco Alfa;
2. Compra a prazo, pagamento em 3 parcelas mensais de terreno – R$ 100,00;
3. Compra à vista de móveis e utensílios – R$ 200,00;
4. Pagamento de 50% da dívida referente à compra do terreno;
5. Compra de mercadorias para revenda por R$ 150,00 a prazo;
6. Obtenção de um empréstimo bancário de curto prazo no valor de R$ 300,00 com depósito no Banco Alfa;
7. Aumento de capital social com um veículo no valor de R$ 80,00;
8. Compra de um caminhão no valor de R$ 170,00 à vista;
9. Pagamento da 1ª parcela do empréstimo no valor R$ 30,00;
10. Compra de mercadorias para revenda no valor de R$ 5,00 à vista;
11. Pagamento de 1/3 da dívida referente ao item 5.

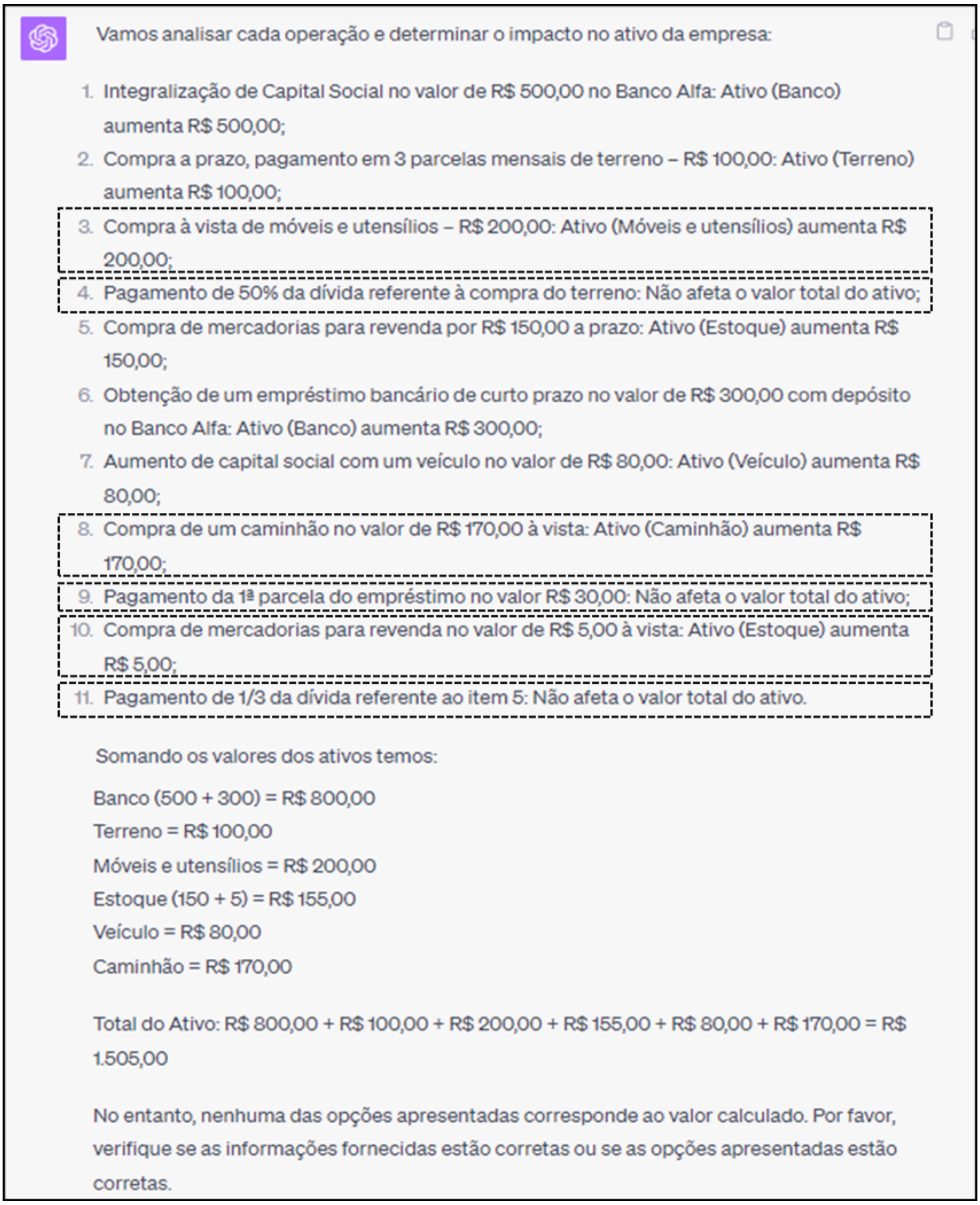
De acordo com as informações apresentadas anteriormente, o valor total do ativo, considerando a NBC TG – Estrutura Conceitual e a NBC TG 26 (R5) – Apresentação das Demonstrações Contábeis, é de:

A) R$ 920,00
B) R$ 1.000,00
C) R$ 1.030,00
D) R$ 1.080,00

Source: authors (2023) and Consulplan (2021)

The response proposed by ChatGPT did not arrive at the correct alternative. The model-generated response can be seen in **Figure 4**.

Figure 4 - ChatGPT Prompt – Example of Response[4]



Vamos analisar cada operação e determinar o impacto no ativo da empresa:

1. Integralização de Capital Social no valor de R$ 500,00 no Banco Alfa: Ativo (Banco) aumenta R$ 500,00;
2. Compra a prazo, pagamento em 3 parcelas mensais de terreno – R$ 100,00: Ativo (Terreno) aumenta R$ 100,00;
3. Compra à vista de móveis e utensílios – R$ 200,00: Ativo (Móveis e utensílios) aumenta R$ 200,00;
4. Pagamento de 50% da dívida referente à compra do terreno: Não afeta o valor total do ativo;
5. Compra de mercadorias para revenda por R$ 150,00 a prazo: Ativo (Estoque) aumenta R$ 150,00;
6. Obtenção de um empréstimo bancário de curto prazo no valor de R$ 300,00 com depósito no Banco Alfa: Ativo (Banco) aumenta R$ 300,00;
7. Aumento de capital social com um veículo no valor de R$ 80,00: Ativo (Veículo) aumenta R$ 80,00;
8. Compra de um caminhão no valor de R$ 170,00 à vista: Ativo (Caminhão) aumenta R$ 170,00;
9. Pagamento da 1ª parcela do empréstimo no valor R$ 30,00: Não afeta o valor total do ativo;
10. Compra de mercadorias para revenda no valor de R$ 5,00 à vista: Ativo (Estoque) aumenta R$ 5,00;
11. Pagamento de 1/3 da dívida referente ao item 5: Não afeta o valor total do ativo.

Somando os valores dos ativos temos:

Banco (500 + 300) = R$ 800,00
Terreno = R$ 100,00
Móveis e utensílios = R$ 200,00
Estoque (150 + 5) = R$ 155,00
Veículo = R$ 80,00
Caminhão = R$ 170,00

Total do Ativo: R$ 800,00 + R$ 100,00 + R$ 200,00 + R$ 155,00 + R$ 80,00 + R$ 170,00 = R$ 1.505,00

No entanto, nenhuma das opções apresentadas corresponde ao valor calculado. Por favor, verifique se as informações fornecidas estão corretas ou se as opções apresentadas estão corretas.

Source: ChatGPT 4 (2023) and authors' emphasis (2023)

The errors made by the model were highlighted with a dashed rectangle. Understanding how NLP models function helps understand the potential reasons for the incorrect responses provided by ChatGPT. It seems the model couldn't comprehend the cash implications of certain operations (cash purchases and loan or financing payments), although it did understand their implications on individual items (Land, Furniture, and Utensils, etc.). In this regard, the model seems to have succeeded: it understood the individual effects of these transactions on each of these accounts, except for the bank account.

It's possible that if the exam question had explicitly stated that the payments were made via bank transfers, for example, the response would have been different and more accurate. Although an accounting student would have some ease in understanding that, given no other options, (the company had no cash) the payments and receipts should occur via the bank, for an NLP model, context is crucial, and the lack of appropriate context can impair understanding and consequently the generated responses.

Another important point is that ChatGPT is not connected to the internet, and the dataset it accesses is an offline database. Therefore, it's not possible to know if the model was trained with the Brazilian Accounting Standards, although it can answer questions about them. It's unknown whether the answers ChatGPT provides are based on the standards or on sites citing the standards.

Moreover, NLP models like this depend on intense human effort in the construction stage. This effort is often manual and repetitive, requiring the analysis of hundreds of thousands of inputs (prompts) and outputs (expected results). Also, it's uncertain how well this model was validated by professionals in business, accounting, or auditing fields. Ultimately, it's crucial to understand that NLP models like this may produce what's called hallucinations — a response delivered with confidence but completely outside the desired reality.

It's worth mentioning that, by the researchers' choice, only one initial prompt was provided (guidance on what the model should do), and subsequently, the exam questions were presented. In no instance was the model corrected by the researchers. It's possible that if the researchers had opted to make corrections, improved prompts throughout the process, and instructed the model with some examples, the performance might have been superior to that presented in this study.

Some implications for accounting (considering that the Proficiency Exam aims to measure professionals' technical capacity and qualify them for the profession) that can be drawn from the results of this research are:

i) NLP models (like ChatGPT) show the ability to assist in some accounting routines relatively effectively and efficiently. Given the high level of accuracy in ethical aspect questions, for example, a professional could use such a tool to verify if a particular situation might be generating an ethical transgression;

ii) NLP models can aid in communication processes with clients and suppliers (improving texts and correcting grammatical issues) and in reviewing accounting reports (internal audit reports, for example).

However, all limitations inherent in these models must be taken into account in their use. Therefore, the results demonstrate that for now, models like NLP are more capable of helping professionals, acting as an assistant, than replacing them in their daily activities. Accounting professionals will need to understand the potentials and limitations of these models to avoid mistakes in their use. The same implications found in this research that apply to the accounting field should also be viewed in other areas of knowledge.

5. Conclusions

The aim of this research was to identify and assess the performance of the ChatGPT tool, GPT-4 model, in the Accounting Proficiency Exam, using this exam as a proxy for accounting activity issues.

The results achieved after applying ChatGPT in four editions of the Accounting Proficiency Exam were significant. The tool demonstrated an average score of 71%, ranging from 64 to 78%, which was sufficient for professional accounting qualification, as the current legislation demands a minimum score of 50%. The areas in which ChatGPT showed higher proficiency were Applied Portuguese Language, Notions of Law and Applied Legislation, and Professional Ethics and Legislation, with accuracy rates of 88, 83, and 81%, respectively. Conversely, topics like Accounting Principles and Brazilian Accounting Standards, Controllership, and Accounting Theory proved to be more challenging for the tool, with scores of 31, 50, and 69%, respectively.

This research contributes theoretically by being a pioneer in reflecting on the impact of NLP models in Brazilian accounting. It can be inferred that while ChatGPT is competent in broader areas like Portuguese Language and Law, it faces challenges in more specific and technical areas of accounting. The question that arises is whether these models will act more as an assistant or replace the professional entirely? Based on the results, it's possible that AI will first establish itself as an assistant in broader areas before advancing into more specific accounting domains.

From a practical perspective, ChatGPT's ability to efficiently solve questions highlights the possibility of its application as an auxiliary tool in accounting education and practice. Furthermore, it serves as a warning for professionals and educational institutions about the need for continuous improvement and the integration of technology into the curriculum and professional practice. However, it's important to note that although ChatGPT succeeded in a significant portion of the questions, there is still room for improvement.

Moreover, it's crucial to consider that the model is a support tool and should not replace human study and knowledge because, due to its nature of continuous learning, it may exhibit unsatisfactory performance even in activities where it currently shows significant accuracy. It's recommended for future studies to investigate if specific model training (prompt engineering) could enhance the tool's performance in areas where it currently faces challenges.

References

Accounting Today (2023), "We had ChatGPT take the CPA exam — and it failed". https://www.accountingtoday.com/news/we-ran-the-cpa-exam-through-chatgpt-and-it-failed-miserably

AICPA (2019), "A CPA's Introduction to AI: From Algorithms to Deep Learning, What You Need to Know". Recuperado de https://us.aicpa.org/content/dam/aicpa/interestareas/frc/assuranceadvisoryservices/downloadabledocuments/56175896-cpas-introduction-to-ai-from-algorithms.pdf

Alves, M. A.; Silva, C.A.T.; Bonfim, M. P. (2023), "ChatGPT e Integridade Acadêmica: Percepção dos Alunos de Contabilidade sobre a Honestidade do Uso do ChatGPT", In Congresso UnB de Contabilidade e Governança, 9.

Amaral, J.V.; Guerreiro, R.; Russo, P. T.; Mucci, D. M. (2023), "Indústria 4.0: Características e potenciais impactos no ambiente interno das empresas", In USP International Conference on Accounting, 23.

Baidoo-Anu, D.; Owusu Ansah, L. (2023), "Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning", doi: 10.2139/ssrn.4337484

Bommarito, J.; Bommarito, M.; Katz, D. M.; Katz, J. (2023), "Gpt as knowledge worker: A zero-shot evaluation of (ai) CPA capabilities". arXiv, preprint, doi: 10.48550/arXiv.2301.04408

Bordt, S.; von Luxburg, U. (2023), "Chatgpt participates in a computer science exam", arXiv, preprint, doi: 10.48550/arXiv.2303.09461

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; ... Amodei, D. (2020), "Language models are few-shot learners", Advances in neural information processing systems, Vol. 33, pp. 1877-1901.

CFC. Conselho Federal de Contabilidade (2023), "Exame de Suficiência". Recuperado de https://cfc.org.br/category/exame-de-suficiencia-anteriores/.

ChatGPT 4 - IA aberta (2023). "ChatGPT" (versão de 15 de maio). https://chat.openai.com/chat

Choi, J. H.; Hickman, K. E.; Monahan, A.; Schwarcz, D. (2023), "Chatgpt goes to law school", doi: 10.2139/ssrn.4335905

Chui, M.; Manyika, J.; Miremadi, M. (2016), "Where machines could replace humans—and where they can't (yet)", McKinsey Quarterly. Recuperado de https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/where-machines-could-replace-humans-and-where-they-cant-yet

Consulplan, Exame 2021/1 (2021), "Exame de Suficiência como Requisito para Obtenção de Registro Profissional em Conselho Regional de Contabilidade (CRC) 02/2021". Recuperado de https://cfc.org.br/exame-de-suficiencia-anteriores/1o-exame-de-suficiencia-de-2021/

Davenport, T. H.; Kirby, J. (2016), "Only humans need apply: Winners and losers in the age of smart machines", Harper Business.

Else, H. (2023), "Abstracts written by ChatGPT fool scientists", Nature, Vol. 613, Num. 7944, pp. 423-423, doi: 10.1038/d41586-023-00056-7

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, Vol. 9, Num. 1, e45312, doi: 10.2196/45312.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). "Design Science in Information Systems Research". MIS Quarterly, Vol. 28, Num. 11, pp. 75–105. Doi: 10.2307/25148625

Jurafsky, D.; Martin, J. H. (2023), "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", Third Edition

Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; ... Kasneci, G. (2023), "ChatGPT for good? On opportunities and challenges of large language models for education", Learning and Individual Differences, Vol. 103, 102274, doi: 10.35542/osf.io/5er8f

Katz, D. M.; Bommarito, M. J.; Gao, S.; & Arredondo, P. (2023), "GPT-4 Passes the Bar Exam" (March 15, 2023). doi: 10.2139/ssrn.4389233

Kenton, J. D. M. W. C.; Toutanova, L. K. (2019, June), "Bert: Pre-training of deep bidirectional transformers for language understanding", In Proceedings of naacL-HLT (Vol. 1, p. 2). doi: 10.18653/v1/N19-1423

Kogan, A.; Sudit, E. F.; Vasarhelyi, M. A. (2017), "Continuous online auditing: A program of research", Journal of Information Systems, Vol. 31, Num. 1, pp. 61-90. doi: 10.2308/jis.1999.13.2.87

Lara, J. E. (2023), "Reflexões editoriais sobre a evolução da ciência e a contribuição do ChatGPT", Revista Gestão & Tecnologia, Vol. 23, Num. 1, pp. 1-3. doi: 10.20397/2177-6652/2023.v23i1.2548

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. (2023), "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing", ACM Computing Surveys, Vol. 55, Num. 9, pp. 1-35. doi: 10.1145/3560815

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; … Tang, J. (2021), "GPT understands, too". doi: 10.48550/arXiv.2103.10385

Loughran, T.; McDonald, B. (2016), "Textual analysis in accounting and finance: A survey", Journal of Accounting Research, Vol. 54, Num. 4, pp. 1187-1230. doi: 10.1111/1475-679X.12123

Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. (2018), "Statistical and machine learning forecasting methods: Concerns and ways forward", PloS one, Vol. 13, Num. 3, e0194889. doi: 10.1371/journal.pone.0194889

Mancebo, V.O.C.; Mucci, D. M. (2023), "Impactos das Tecnologias Digitais nas Atividades Desempenhadas pela Controladoria", In USP International Conference on Accounting, 23.

Martínez, E. (2023), "Reavaliando o desempenho do GPT-4 no exame de barra" (8 de maio de 2023). Documento de Trabalho LPP No. 2-2023, doi: 10.2139/ssrn.4441311

Miranda, C. D. S.; Araújo, A. M. P. D.; Miranda, R. A. D. M. (2017), "O exame de suficiência em contabilidade: uma avaliação sob a perspectiva dos pesquisadores", Revista Ambiente Contábil, Vol. 9, Num. 2, pp. 158-178. doi: 10.16930/2237-766220202952

Mujiono, M. N. (2021), "The shifting role of accountants in the era of digital disruption", International Journal of Multidisciplinary: Applied Business and Education Research, Vol. 2, Num. 11, pp. 1259-1274. doi: 10.11594/10.11594/ijmaber.02.11.18

OpenAI (2023). Recuperado de dttps://openai.com/research/gpt-4, Acesso em 12, abril 2023

Pardos, Z. A.; Bhandari, S. (2023), "Learning gain differences between ChatGPT and human tutor generated algebra hints", arXiv preprint. doi: 10.48550/arXiv.2302.06871

Qadir, J. (2022), "Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education". Recuperado de https://www.techrxiv.org/articles/preprint/Engineering_Education_in_the_Era_of_ChatGPT_Promise_and_Pitfalls_of_Generative_AI_for_Education/21789434

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. (2018), "Improving language understanding by generative pre-training".

Subramani, M.; Jaleel, I.; Krishna Mohan, S. (2023), "Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS", Advances in Physiology Education, Vol. 47, Num. 2, pp. 270-271. doi: 10.1152/advan.00036.2023

Vasarhelyi, M. A.; Kogan, A.; Tuttle, B. M. (2015), "Big data in accounting: An overview", Accounting Horizons, Vol. 29, Num. 2, pp. 381-396. doi: 10.2308/acch-51071

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; ... Polosukhin, I. (2017), "Attention is all you need", Advances in neural information processing systems, Vol. 30.

Notes

2. You must act as an Accountant who knows Brazilian Accounting Standards, Cost Accounting, Management Accounting, Tax (related to Brazilian tax aspects) and all other aspects that surround Brazilian accounting. Next, I'm going to give you questions about accounting and I need you to answer them for me. All questions are multiple choice (A, B, C, D). Only a single alternative is correct. For each question, a brief explanation of how you arrived at the result should be given. Furthermore, I ask you to identify which area of accounting that question can be framed in (Auditing, Accounting Applied to the Public Sector, Cost Accounting, General Accounting, Management Accounting, Controllership, Legislation and Professional Ethics, Applied Portuguese Language, Financial Mathematics and Statistics, Concepts of Law and Applied Legislation, Forensic Accounting, Accounting Principles and Brazilian Accounting Standards, Accounting Theory).
Are you ready?

3. A certain business company presented, in May 2020, the following operations:
1. Share Capital Contribution in the amount of R$ 500.00 at Alfa Bank;
2. Deferred payment purchase of land, with payment in 3 monthly installments – R$ 100.00;
3. Cash purchase of furniture and utensils – R$ 200.00;
4. Payment of 50% of the debt related to the land purchase;
5. Purchase of goods for resale for R$ 150.00 in installments;
6. Obtaining a short-term bank loan in the amount of R$300.00 with a deposit in Alfa Bank;
7. Increase in share capital with a vehicle valued at R$80.00;
8. Purchase of a truck worth R$ 170.00 in cash;
9. Payment of the 1st installment of the loan in the amount of R$ 30.00;
10. Purchase of goods for resale worth R$5.00 in cash;
11. Payment of one-third of the debt referring to item 5.
According to the information presented previously, the total value of the asset, considering NBC TG – Conceptual Structure and NBC TG 26 (R5) – Presentation of Financial Statements, is:
A) R$ 920.00
B) R$ 1,000.00
C) R$ 1,030.00
D) R$ 1,080.00

4. Let's analyze each operation and determine the impact on the company's assets:
1. Share Capital Contribution in the amount of R$ 500.00 at Alfa Bank: Assets (Bank) increase by R$500.00;
2. Deferred payment purchase of land, with payment in 3 monthly installments – R$ 100.00: Assets (Land) increase by R$ 100.00;
3. Cash purchase of furniture and utensils – R$ 200.00: Assets (Furniture and utensils) increase by R$200.00 ;
4. Payment of 50% of the debt related to the land purchase: Does not affect the total value of the asset;
5.Purchase of goods for resale for R$ 150.00 in installments: Assets (Stock) increase by R$150.00;
6. Obtaining a short-term bank loan in the amount of R$300.00 with a deposit in Alfa Bank: Assets (Bank) increase by R$300.00;
7. Increase in share capital with a vehicle valued at R$80.00: Assets (Vehicle) increase by R$80.00;
8.Purchase of a truck worth R$ 170.00 in cash: Assets (Truck) increase by R$170.00;
9.Payment of the 1st installment of the loan in the amount of R$ 30.00: Does not affect the total value of the asset;
10.Purchase of goods for resale worth R$5.00 in cash: Assets (Stock) increase by R$5.00;
11. Payment of one-third of the debt referring to item 5: Does not affect the total value of the asset.
Adding the asset values we have:
Bank (500 + 300) = R$800.00 Land = R$ 100.00 Furniture and utensils = R$ 200.00 Stock (150 + 5) = R$ 155.00 Vehicle = R$80.00 Truck = R$ 170.00
Total Assets: R$800.00 + R$100.00 + R$200.00 + R$155.00 + R$80.00 + R$170.00 = R$1,505.00
However, none of the options presented correspond to the calculated value. Please check that the information provided is correct or that the options presented are correct