

Evolução da IA em resolver problemas contábeis: uma comparação entre o GPT4 e GPT4o na resolução do Exame de Suficiência

Marcelo Machado de Freitas

Universidade Federal de Santa Catarina - UFSC, Brasil

Jonatas Dutra Sallaberry

Universidade do Contestado - UNC, Brasil

Thiago Bruno de Jesus Silva

Universidade Federal do Recôncavo da Bahia - UFRB, Brasil

Fabricia Silva da Rosa

Universidade Federal de Santa Catarina - UFSC, Brasil

Evolución de la IA en la resolución de problemas contables: Comparación entre GPT4 y GPT4o en el Examen de Competencia Contable

La investigación tuvo como objetivo evaluar el rendimiento del modelo GPT-4o en comparación con el modelo GPT-4 en la resolución de preguntas del Examen de Competencia Contable. Este estudio se fundamenta en el concepto de procesamiento del lenguaje natural (PLN) de Brown et al. (2020). La investigación se realizó utilizando la metodología *design science*, que busca construir y/o evaluar diferentes artefactos tecnológicos aplicando las preguntas del Examen de Competencia al GPT-4o, utilizando ChatGPT de OpenAI. Si bien las estadísticas reales del Examen de Competencia muestran que solo una parte de los contadores lo aprueba, los resultados de la inteligencia artificial (IA) mostraron que las cuatro ediciones evaluadas aprobaron con al menos 64% de éxito. En general, en la muestra analizada, el modelo de IA GPT-4o alcanzó 77% de precisión, en comparación con el 71% del modelo GPT-4; así, se alcanzó un 84% de precisión en los dos últimos exámenes con el modelo más reciente. Sin embargo, en algunas preguntas



<https://doi.org/10.18800/contabilidad.2025ESP.005>

Contabilidad y Negocios 20 (esp.) 2025, pp. 109-126 / e-ISSN 2221-

que el modelo GPT-4 había respondido correctamente previamente, el modelo más reciente terminó respondiendo incorrectamente o de forma diferente a la esperada por los desarrolladores. Estos resultados contribuyen a la literatura que estudia el uso de la IA en contabilidad, en particular, en la aplicación de modelos de procesamiento del lenguaje natural y modelos de lenguaje extenso (LLMs).

Palabras clave: uso de tecnologías, ChatGPT4o, modelos de lenguaje extensos (LLMs)

Evolution of AI in solving accounting problems: A comparison between GPT4 and GPT4o in solving the Accounting Proficiency Exam

The research aimed to evaluate the performance of the GPT-4o model compared to the GPT-4 model in solving questions on the Accounting Proficiency Exam. This study is grounded in the concept of natural language processing (PLN), as discussed by Brown et al. (2020). The research was conducted using the design science methodology, which aims to build and/or evaluate different technological artifacts by applying the proficiency exam questions to GPT-4o, using OpenAI's ChatGPT. While actual statistics for the Proficiency Exam show that only a portion of accountants pass, the results of artificial intelligence (AI) showed that all four editions evaluated passed with at least a 64% success rate. Overall, across the sample analyzed, the GPT-4o AI model achieved 77% accuracy compared to 71% for the GPT-4 model, achieving 84% accuracy in the last two exams using the most recent model. However, on some questions that the GPT-4 model had previously answered correctly, the more recent model ended up answering incorrectly or differently than expected by the question developers. These results contribute to the literature studying the use of AI in accounting, particularly the application of natural language processing models and large language models (LLMs).

Keywords: use of technologies, ChatGPT4o, large language models (LLMs)

Evolução da IA em resolver problemas contábeis: uma comparação entre o GPT4 e GPT4o na resolução do Exame de Suficiência

A pesquisa teve por objetivo avaliar o desempenho do modelo GPT-4o em relação ao modelo GPT-4 na resolução de questões do Exame de Suficiência em Contabilidade. Esse estudo se baseia no conceito de processamento de linguagem natural (PLN) de Brown et al. (2020). A pesquisa foi operacionalizada com a metodologia de *design science*, que visa construir e/ou avaliar diferentes artefatos tecnológicos, mediante a aplicação das questões do Exame de Suficiência no GPT-4o, por meio do ChatGPT da OpenAI. Enquanto as estatísticas reais do Exame de Suficiência demonstram a aprovação de apenas uma parcela dos contabilistas, os resultados da inteligência artificial (IA) foram de aprovação em todas as quatro edições avaliadas com pelo menos 64% de aproveitamento. No âmbito geral da amostra analisada, o modelo de IA do GPT-4o apresentou 77% de aproveitamento de respostas frente a 71% do modelo GPT-4, alcançando uma acurácia de 84% nos últimos dois exames pelo modelo mais recente. Ainda assim, em algumas questões que o modelo GPT-4 havia acertado, esse modelo mais atual acabou respondendo de forma equivocada, ou diferente do previsto pelos elaboradores das questões. Esses resultados contribuem para a literatura

que estuda o uso de IA dentro do campo da contabilidade, em especial a aplicação de modelos de linguagem de processamento natural e *large language models* (LLMs).

Palavras-chave: uso de tecnologias, ChatGPT4o, large language models (LLMs)

1. INTRODUÇÃO

Modelos de processamento de linguagem natural (PLN) podem realizar análise de sentimentos, responder perguntas e classificar textos (Essel et al., 2024). Esses modelos permitem desde a automação de tarefas repetitivas até o aprimoramento dos processos de tomada de decisão (Davenport & Kirby, 2016), além de possibilitar a análise de dados financeiros, auxiliando em previsões e decisões estratégicas (Makridakis et al., 2018). Especificamente, o ChatGPT é capaz de gerar respostas em formato de conversa semelhantes às humanas (Radford et al., 2018), pois foi treinado com uma vasta quantidade de dados textuais, permitindo que entenda e gere textos com alta precisão (Kasneci et al., 2023). A IA pode ser ainda, uma ferramenta valiosa no ensino contábil, proporcionando *feedback* instantâneo e orientações personalizadas aos estudantes (Albuquerque & Gomes do Santos, 2024; Baidoo-Anu & Owusu Ansah, 2023).

No entanto, a rápida evolução dessas tecnologias impressiona e, ao mesmo tempo, serve de alerta para a profissão contábil. Por um lado, existem preocupações com os desafios relacionados à eficiência dos modelos e, por outro, com o impacto na formação de contadores e na continuidade da profissão como a conhecemos. É importante reconhecer que os modelos de PLN ainda enfrentam desafios, como a geração de outputs falsos ou tóxicos e limitações no raciocínio abstrato (Wu et al., 2023).

Estudos anteriores revelam que modelos de PLN, como o GPT-4, podem ser utilizados para facilitar a comunicação entre alunos e instrutores (Pardos & Bhandari, 2024; Qadir, 2023) e analisar os exames de suficiência de contabilidade (Albuquerque & Gomes do Santos, 2024; Atanasovski et al., 2023; Eulerich et al., 2024; Freitas et al., 2024; Masuda et al., 2023). Os resultados desses estudos revelam as capacidades da inteligência artificial (IA) em satisfazer problemas teóricos e práticos, que podem ser respondidos em termos descritivos, do mundo contábil (Freitas et al., 2024). No entanto, há um contraditório: por um lado, teme-se que o alto desempenho do ChatGPT possa ser prejudicial aos setores de contabilidade e auditoria (Eulerich et al., 2024), e por outro, estudos de Albuquerque e Gomes do Santos (2024), e Freitas et al. (2024) identificaram que a ferramenta é menos eficiente em questões que exigem maior julgamento profissional, demonstrando falta de precisão em questões mais

complexas nos exames de proficiência em contabilidade (Albuquerque & Gomes do Santos, 2024).

Essa ambiguidade destaca a necessidade de ampliar as análises sobre a contribuição do uso de modelos de PLN, como o GPT4 e o GPT4o, para a profissão contábil, uma vez que estudos anteriores também revelam que esta IA tem a capacidade de evoluir rapidamente, melhorando suas análises e suas respostas em exames de proficiência (Freitas et al., 2024). Dado esse contexto, o estudo objetiva avaliar o desempenho do modelo GPT-4o em relação ao modelo GPT-4 na resolução de questões do Exame de Suficiência em Contabilidade, como *proxy* de problemas na atividade contábil.

Decidiu-se empregar as questões do Exame de Suficiência como uma *proxy* os problemas na atividade contábil, ainda que questões teóricas ou práticas do conteúdo contábil não refletem completamente a capacidade de executar tarefas avançadas de um profissional contábil. Todavia, essa escolha é justificada pela importância do exame, uma vez que sua aprovação é uma exigência legal para o exercício da profissão contábil no Brasil.

Este estudo contribui ao analisar a contribuição da IA com a área de contabilidade, e acompanhar a integração da tecnologia no ensino e na prática contábil, alertando para a importância de considerar a IA como uma ferramenta de apoio e não um substituto ao conhecimento humano. Também serve para investigar como os treinamentos específicos no modelo melhoram seu desempenho em áreas desafiadoras.

2. REFERENCIAL TEÓRICO

A IA e os modelos de PLN têm o potencial de revolucionar diversos aspectos da contabilidade, desde a automação de tarefas rotineiras até a melhoria dos processos de tomada de decisão (Davenport & Kirby, 2016). De acordo com Chui et al. (2016), uma parte considerável das atividades realizadas por profissionais financeiros pode ser automatizada em um futuro próximo, especialmente aquelas ligadas à coleta e ao processamento de dados.

O alcance das aplicações de IA vai além, abrangendo a solução de problemas, PLN, reconhecimento de fala, análise de imagens, programação automática e robótica (Guerreiro et al., 2023). Essas tecnologias não apenas armazenam e processam dados com maior eficácia e eficiência do que os humanos, mas também permitem a substituição de trabalhadores por máquinas em certos processos, além de possibilitar a realocação de funcionários para funções mais estratégicas e críticas nos negócios (Mancebo & Mucci, 2023).

Os modelos de PLN podem extrair informações de dados não estruturados, como demonstrativos financeiros, relatórios anuais e notas de auditores (Loughran & McDonald, 2016). Integrando dados históricos e informações financeiras em tempo real, esses algoritmos de aprendizado de máquina são capazes de identificar tendências e padrões, auxiliando analistas humanos. Isso conduz a previsões mais precisas e a decisões estratégicas mais fundamentadas (Makridakis et al., 2018). Essa capacidade de análise aprimorada pode, portanto, modificar a maneira como as empresas de contabilidade operam e tomam decisões estratégicas.

O ChatGPT pode gerar relatórios de previsão financeira e sugestões de tomada de decisão com base nos dados fornecidos, identificar e avaliar riscos financeiros e auxiliar o pessoal da contabilidade na condução de análises financeiras inteligentes. Sua aplicação específica na gestão financeira poderá contribuir para a transformação digital das empresas. (Fu et al., 2024). Foram observados estudos com foco em dois pontos relacionados ao ChatGPT em contabilidade e finanças, de acordo ao estudo bibliométrico por Dong et al. (2024). O primeiro foi centrado na utilidade do ChatGPT como ferramenta de pesquisa, com o intuito de aproveitar a capacidade como classificação, resumo e geração de texto. O segundo, centraram nas implicações da adoção à profissão, bem como para diversas organizações e setores.

Modelos de PLN, como o GPT-4 e o GPT-4o, podem ser aplicados para melhorar a comunicação entre alunos e instrutores, oferecendo *feedback* imediato e explicações sobre conceitos complexos. Essa eficácia já foi comprovada em outras áreas do conhecimento, como a matemática (Pardos & Bhandari, 2024) e a engenharia (Qadir, 2023). O ChatGPT já foi utilizado para analisar os exames de suficiência para advogados e engenharia, por exemplo (Katz et al, 2024; Martínez, 2024; Pursnani et al., 2023). Em contabilidade foram encontrados alguns estudos com a mesma finalidade em contextos distintos, como Albuquerque e Gomes do Santos (2024), Atanasovski et al. (2023), Eulerich et al. (2024), Freitas et al. (2024), Masuda et al. (2023), e Oliveira e Khatib (2024).

Os resultados dos estudos sugerem a importância de considerar o Exame de Suficiência em Ciências Contábeis como uma medida representativa dos conhecimentos necessários para o exercício da profissão contábil, oferecendo uma visão inicial das capacidades da IA em atender a questões teóricas e práticas do mundo contábil, que podem ser descritas de forma detalhada (Freitas et al., 2024).

Eulerich et al. (2024) apontam que o alto desempenho do ChatGPT indica que a ferramenta pode, de fato, impactar os setores de contabilidade e auditoria. Por outro lado, Albuquerque e Gomes do Santos (2024), e Freitas et al. (2024) observaram que

a ferramenta não foi uniformemente precisa em todas as questões, especialmente aquelas que exigiam julgamento profissional mais aprofundado.

Esses resultados ressaltam que, embora o ChatGPT possa ser uma ferramenta útil para aumentar a eficiência dos profissionais, ele não está, pelo menos por ora, em posição de substituir o profissional humano. A utilidade potencial do ChatGPT também se estende ao ambiente educacional, servindo como um recurso em salas de aula para incentivar o pensamento crítico dos alunos e facilitar discussões orientadas pelos professores (Albuquerque & Gomes do Santos, 2024). Esta aplicação pode preparar os estudantes para uma utilização mais eficaz da tecnologia em suas futuras práticas profissionais.

O Exame de Suficiência, estabelecido pelo Conselho Federal de Contabilidade (CFC) através da *Resolução CFC nº 853/1999* (CFC, 1999) e consolidado pela *Lei nº 12.249, de 11 de junho de 2010* (Brasil, 2010), é uma avaliação destinada a verificar se os bacharéis em Ciências Contábeis possuem os conhecimentos necessários para exercer a profissão com um nível básico de precisão. Inspirado no modelo americano *certified public accountant* (CPA), o exame busca garantir um padrão mínimo de competência técnica para contadores no Brasil. O exame é realizado semestralmente e consiste em questões objetivas e dissertativas que cobrem os conteúdos fundamentais do curso de bacharelado em Ciências Contábeis (Miranda et al., 2017).

Embora seja amplamente reconhecido como um mecanismo de valorização profissional e incentivo à educação contínua (Lima et al., 2012; Miranda et al., 2017; Nogueira et al., 2013; Terres et al., 2011), o exame enfrenta críticas quanto à sua capacidade de refletir adequadamente a preparação dos estudantes. As discrepâncias entre o conteúdo do exame e as exigências globais e profissionais, além da falta de alinhamento com o currículo sugerido pela Organização das Nações Unidas (ONU) e as demandas do mercado (Oliveira Neto & Kurotori, 2009), ressaltam as limitações do formato atual. Pesquisas apontam que o desempenho no Exame de Suficiência não está diretamente relacionado à carga horária das disciplinas, embora haja uma correlação significativa e positiva com o desempenho no Exame Nacional de Desempenho dos Estudantes (ENADE) (Sena & Sallaberry, 2021).

A análise dos resultados das edições mais recentes do Exame de Suficiência revela uma realidade preocupante: a baixa taxa de aprovação entre os candidatos. No exame de 2022-2, por exemplo, dos 44.085 inscritos, 35.984 compareceram e apenas 7.595 foram aprovados, representando 21,11% dos presentes (CFC, 2023). Esses dados indicam que muitos graduados em Ciências Contábeis enfrentam dificuldades para ingressar na profissão. Além disso, entre os candidatos que prestaram o exame pela

primeira vez, somente 38% foram aprovados, enquanto 62% eram candidatos que já haviam sido reprovados em tentativas anteriores (CFC, 2022).

Os resultados também expõem disparidades entre diferentes grupos de conteúdo. Essa tabela detalha a distribuição de questões por área do conhecimento e o desempenho percentual dos candidatos de 2018 a 2022, além de apresentar a média de desempenho para o período mencionado (Miranda et al., 2017).

3. PROCEDIMENTOS METODOLÓGICOS

A pesquisa foi operacionalizada utilizando a metodologia *design science*, que visa construir e avaliar diferentes artefatos (tecnológicos ou não tecnológicos) dentro de um campo específico do conhecimento (Hevner et al., 2004). Especificamente, a pesquisa testa a aplicabilidade de modelos de PLN e seus *large language models* (LLMs) dentro do ambiente contábil. Hevner et al. (2004) explicam que o paradigma da *design science* busca expandir os limites das capacidades humanas e organizacionais por meio da criação e avaliação de artefatos inovadores.

Primeiramente foi necessário selecionar uma *proxy* de problemas na atividade contábil onde a pesquisa pudesse ser operacionalizada. Como a aprovação no Exame de Suficiência do CFC é condição necessária para que os profissionais da área possuam atuar efetivamente nas funções de contabilidade, escolheu-se as questões desse exame como uma *proxy* para esses problemas.

As questões utilizadas para serem respondidas pelo modelo do GPT-4 e GPT-4o foram extraídas dos Exames de Suficiência dos anos de 2021 e 2022 (Prova Branca). As questões foram coletadas, tabuladas com as respectivas alternativas de respostas, e posteriormente inseridas manualmente (uma a uma), dentro da interface do ChatGPT.

Freitas et al. (2024) desenvolveram e validaram com especialistas uma indicação condicional específica para ser usada como *prompt* inicial na interação com a tecnologia, considerando os modelos de perguntas e os objetivos da pesquisa. Essa questão foi então inserida no início da conversa com o ChatGPT, modelo GPT-4 e GPT-4o.

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. Está pronto? (OpenAI, 15 de julho de 2024).

Cada uma das 50 questões, de cada uma das provas, do CFC foi inserida no chat, uma a uma, sem qualquer novo *prompt* ao longo da execução. Embora *prompts* adicionais pudessesem ter sido incluídos para tentar melhorar a qualidade das respostas, optou-se por manter um *prompt* que somente delimitasse os conteúdos a serem abordados, procurando testar a capacidade do modelo de entender adequadamente o contexto de cada questão e respondê-la. No caso de questões que apresentavam tabelas, os dados foram extraídos e apresentados sequencialmente por linhas.

As questões do exame do CFC não foram inseridas todas em uma única sessão. Dessa forma, elas foram inseridas no ChatGPT entre os dias 10 de agosto e 27 de agosto de 2024. Toda vez que uma nova sessão era iniciada, o *prompt* inicial era novamente inserido para que o modelo do GPT sempre tivesse as mesmas orientações.

Após a resolução das questões, os resultados foram organizados em planilhas, onde as respostas fornecidas pelo ChatGPT foram registradas e avaliadas em relação ao gabarito oficial de cada exame. No caso de questões que foram anuladas pelo CFC e pela banca examinadora, as respostas do ChatGPT foram consideradas corretas. Para alinhar com os objetivos da pesquisa, os autores classificaram as questões com base nos conteúdos e quantitativos especificados pela banca organizadora, facilitando a análise por conteúdo.

As análises das respostas do ChatGPT foram realizadas tanto quantitativamente - para destacar o número de acertos por prova e por área - quanto qualitativamente. O objetivo da análise qualitativa era compreender como as respostas foram formuladas e identificar os principais erros cometidos pelo modelo, proporcionando *insights* sobre as capacidades e limitações do sistema em contextos de teste específicos.

4. RESULTADOS E DISCUSSÕES

Percebe-se que o modelo GPT-4o superou, em média, o modelo GPT-4, com um aumento percentual médio de acerto de 8,5%. Considerando que o GPT-4 já havia apresentado um resultado relativamente satisfatório, esse incremento coloca o GPT-4o com uma nota muito próxima (ou até maior) das alcançadas pelos melhores resultados daqueles que realizaram a prova. Por exemplo, no 2022-2, as notas máximas nas regiões brasileiras foram: centro-oeste (43 pontos), nordeste (44 pontos), norte (38 pontos), sudeste (44 pontos) e sul (45 pontos), enquanto o modelo GPT-4o alcançou um

resultado de 42 acertos. Nesse exame, inclusive, o modelo teve um desempenho 31% maior que seu modelo antecessor.

Tabela 1. Resultados da pesquisa

Especialidade	Total	GPT4		GPT40		GPT4		GPT40													
		GPT4		GPT40		GPT4		GPT40		GPT4		GPT40		GPT4		GPT40		GPT4		GPT40	
		Acertos 21-1	(%) Aprov. 21-1	Acertos 21-1	(%) Aprov. 21-1	Acertos 21-2	(%) Aprov. 21-2	Acertos 21-2	(%) Aprov. 21-2	Acertos 22-1	(%) Aprov. 22-1	Acertos 22-1	(%) Aprov. 22-1	Acertos 22-2	(%) Aprov. 22-2	Acertos 22-2	(%) Aprov. 22-2	Acertos Geral	(%) Aprov. Geral	Acertos Geral	(%) Aprov. Geral
CG	17	11	65	11	65	13	76	11	65	14	82	16	94	10	59	13	76	48	71	51	75
CC	4	4	100	4	100	3	75	4	100	3	75	3	75	2	50	3	75	12	75	14	88
CSP	3	2	67	3	100	2	67	3	100	3	100	3	100	2	67	3	100	9	75	12	100
CGR	2	1	50	0	0	2	100	2	100	2	100	2	100	1	50	2	100	6	75	6	75
CON	1	1	100	1	100	0	0	0	0	1	100	1	100	0	0	0	0	2	50	2	50
DIR	3	2	67	2	67	2	67	3	100	3	100	3	100	3	100	2	67	10	83	10	83
MF	2	2	100	2	100	2	100	2	100	0	0	1	50	2	100	2	100	6	75	7	88
TC	4	3	75	1	25	2	50	2	50	3	75	4	100	3	75	4	100	11	69	11	69
LE	4	4	100	4	100	4	100	3	75	3	75	3	75	2	50	4	100	13	81	14	88
PNC	4	2	50	2	50	1	25	2	50	1	25	2	50	1	25	3	75	5	31	9	56
AC	2	0	0	0	0	2	100	2	100	2	100	2	100	2	100	2	100	6	75	6	75
PC	2	1	50	2	100	1	50	1	50	2	100	1	50	2	100	2	100	6	75	6	75
LP	2	1	50	0	0	2	100	2	100	2	100	1	50	2	100	2	100	7	88	5	63
Total	50	34	68	32	64	36	72	37	74	39	78	42	84	32	64	42	84	141	71	153	77

Nota. Abreviaturas: AC = Auditoria Contábil; CC = Contabilidade de Custos; CG = Contabilidade Geral; CGR = Contabilidade Gerencial; CON = Controladoria; CSP = Contabilidade do Setor Público; DIR = Direito; LE = Legislação e Ética; LP = Língua Portuguesa; MF = Matemática Financeira; PC = Perícia Contábil; PN = Princípios e Normas Contábeis; TC = Teoria da Contabilidade.

No âmbito geral da amostra analisada, o modelo GPT-40 apresentou 77% de aproveitamento de respostas frente a 71% do modelo GPT-4. Todavia, nota-se que no Exame de Suficiência de 2021-1, o modelo GPT-40 gerou um resultado inferior ao seu modelo antecessor. Esse resultado é particularmente interessante, pois existem discussões sobre até onde tais modelos de PLN conseguirão avançar, e quais serão os avanços realmente significantes daqui para frente.

Embora os desenvolvedores dos modelos GPT (Open Ai) indiquem que o GPT-40 é mais preciso do que o modelo GPT-4, o resultado obtido em 2021-1 levanta dúvidas

sobre a real melhora desse novo modelo. Todavia, uma característica do modelo GPT-4o é que ele gera respostas consideravelmente mais rápidas do que o modelo GPT-4, e embora isso não tenha sido mensurado de forma metodológica na presente pesquisa, isso foi observado durante a aplicação das questões no ChatGPT. Essa característica ágil pode ser particularmente importante em alguns ambientes de decisão (como por exemplo, em *chatbots*, em que o cliente necessita de uma resposta mais ágil). Porém, no cenário contábil, a precisão das respostas é mais importante do que a agilidade (especialmente quando a variação no tempo de resposta é levemente diferente de um modelo para o outro). Um cenário contábil onde pode ser particularmente importante mais agilidade seria no envio de um e-mail a clientes contendo informações contábeis. Um modelo que possa rapidamente evidenciar inconsistências entre a informação enviada e as informações contábeis poderia ser útil (por exemplo, o envio de informações de determinado cliente por e-mail para um outro cliente).

Quanto aos resultados por área, observou-se que o modelo GPT-4o geralmente superou o GPT-4 em termos de taxa de acertos e consistência em diversas áreas ao longo dos períodos estudados. Especificamente, em Contabilidade Geral (CG), ambos os modelos apresentaram desempenho semelhante no início do período analisado (2021-1 e 2021-2), mas o GPT-4o mostrou melhorias em semestres subsequentes (2022-1 e 2022-2).

Na área de Contabilidade de Custos (CC) o GPT-4o melhorou seu desempenho em duas, das quatro edições analisadas. Além disso, na área de Contabilidade do Setor Público (CSP), o GPT-4o alcançou 100% de sucesso em todas as edições avaliadas, melhorando seu desempenho em relação ao modelo GPT-4 em três edições.

Percebe-se que no aproveitamento geral, houve uma piora somente na especialidade de Língua Portuguesa (LP), o que é algo um pouco surpreendente, já que a priori deveria ser uma das áreas com maior facilidade para modelos de PLN, como os modelos GPT.

Para entender um pouco melhor os erros cometidos pelo GPT-4o em relação ao modelo GPT-4, foi selecionado uma questão na qual o modelo mais recente trouxe um resultado errado, e o modelo antecessor, havia acertado a mesma questão. Uma dessas questões foi a questão 25 do Exame de Suficiência de 2021-1, na qual será apresentada a seguir.

Figura 1. Questão 25: Uma sociedade empresária apresentou as seguintes informações:

Receita de vendas	800.000,00
Custos e despesas variáveis (CDV)	(400.000,00)
Margem de contribuição (MC)	400.000,00
Custos e despesas fixos (CDF)	(150.000,00)
Resultado operacional (RO)	250.000,00

A empresa quer aumentar sua produção e vendas em 25%, passando para 5.000 unidades mensais, mantendo o custo fixo em R\$ 150.000,00 e o preço de venda unitário em R\$ 200,00. Sabendo-se que o custo total (fixo + variável) aumentará de R\$ 550.000,00 para R\$ 650.000,00, qual será o grau de alavancagem operacional?

A) 1,25 B) 1,40 C) **1,60** D) 2,00 (**Resposta correta em negrito**)

As respostas dos modelos GPT-4 e GPT-4o podem ser visualizadas a seguir, na figura 2.

Percebe-se que enquanto o modelo GPT-4 usou a fórmula tradicional do grau de alavancagem operacional (GAO), calculando a variação no resultado operacional (RO) e na receita de vendas e, em seguida, dividindo essas variações o GPT-4o utilizou a fórmula simplificada do GAO, que é a relação direta entre a margem de contribuição (MC) e o RO. É difícil precisar porque o modelo GPT-4o optou por essa solução, mas uma hipótese plausível é porque esse modelo, dado mais de um caminho viável possível, irá por aquele cujo resultado será dado de forma mais rápida e ágil. Outra possibilidade decorre da evolução do modelo, que avança na interpretação de problemas e elaboração de soluções além da simples aplicação de fórmulas.

É importante reforçar que não foram realizadas técnicas de otimização dos *prompts* utilizados (*prompt engineering*), pois a intenção da pesquisa era analisar as respostas apresentadas pelos modelos com a menor interferência possível por parte dos pesquisadores. Acredita-se que os resultados poderiam ser ainda melhores se essas técnicas fossem implementadas.

Para exemplificar a racionalização apresentada no parágrafo anterior, foi realizada uma análise adicional, sendo inserido no *prompt* do ChatGPT, para resolução da questão 25 do Exame de Suficiência uma outra orientação: “Para responder essa questão, você deve aplicar a fórmula GAO = variação percentual do lucro / variação percentual das vendas”. Essa simples orientação adicional fez o modelo acertar a questão.

Figura 2. Estatísticas por grupo de conteúdo

Resposta GPT-4	Resposta GPT-40
<p>Primeiro, precisamos encontrar o novo resultado operacional com o aumento de 25% na produção e vendas.</p>	<p>Para resolver a questão, precisamos calcular o grau de alavancagem operacional (GAO). O GAO é calculado pela fórmula: $GAO = \text{margem de contribuição} / \text{resultado operacional}$</p>
<p>Novas vendas: $800.000,00 \times 1,25 = 1.000.000,00$</p>	<p>Primeiro, vamos calcular a nova receita de vendas e a nova margem de contribuição com o aumento de 25% na produção e vendas.</p>
<p>Novo custo total: 650.000,00 Novo resultado operacional: $1.000.000,00 - 650.000,00 = 350.000,00$</p>	<p>Passo 1: Calcular a nova receita de vendas Receita de vendas original: R\$800.000,00 Aumento de 25%: $800.000,00 \times 1,25 = R\\$1.000.000,00$</p>
<p>Agora, podemos calcular o grau de alavancagem operacional (GAO) usando a seguinte fórmula:</p>	<p>Passo 2: Calcular a nova margem de contribuição Custos e despesas variáveis (CDV) originais: R\$400.000,00 Aumento nos custos totais (fixos + variáveis) de R\$550.000,00 para R\$650.000,00, resultando em um aumento de R\$100.000,00</p>
<p>$GAO = (\Delta\text{resultado operacional} / \text{resultado operacional inicial}) / (\Delta\text{receita de vendas} / \text{receita de vendas inicial})$</p>	<p>Passo 3: Calcular a nova margem de contribuição: Receita de vendas: R\$1.000.000,00 (-) Custos e despesas variáveis: R\$500.000,00 (já que o custo variável aumentou proporcionalmente ao aumento na produção)</p>
<p>$\Delta\text{Resultado operacional} = 350.000,00 - 250.000,00 = 100.000,00$ $\Delta\text{Receita de vendas} = 1.000.000,00 - 800.000,00 = 200.000,00$</p>	<p>Passo 4: Calcular o novo resultado operacional Custos e despesas fixos (CDF): R\$150.000,00</p>
<p>$GAO = (100.000,00 / 250.000,00) / (200.000,00 / 800.000,00)$</p>	<p>Resultado operacional: $RO = MC - CDF$ $RO = R\\$500.000,00 - R\\$150.000,00 = R\\$350.000,00$</p>
<p>$GAO = (0,4) / (0,25) = 1,6$</p>	<p>Passo 4: Calcular o grau de alavancagem operacional (GAO) $GAO = MC / RO$ $GAO = R\\$500.000,00 / R\\$350.000,00$ $GAO \approx 1,43$</p>
<p>Portanto, o grau de alavancagem operacional é:</p>	<p>Portanto, a resposta correta é: B) 1,40</p>

5. CONCLUSÕES

Foi possível avaliar que houve um incremento médio nos acertos realizados pelo modelo mais atual (GPT-4o) em relação ao seu modelo antecessor (GPT-4). A relevância desse incremento precisa ser avaliada dentro do espaço em que ainda existe de melhora, já que nos exames mais recentes (2022), o GPT-4o alcançou uma taxa de acerto de 84% da prova, o que é um resultado bastante expressivo.

Todavia, a qualidade de algumas respostas dadas pelo modelo mais atual pode ser questionada, já que em algumas questões que o modelo GPT-4 havia acertado, esse modelo mais atual acabou respondendo de forma equivocada, ou diferente do previsto pelos elaboradores das questões. O processamento de soluções parte de algoritmos estabelecidos para entender os questionamentos e apresentar respostas, que poderiam ser de distintas formas, mas devido a parâmetros técnicos precisam se enquadrar entre as respostas disponíveis, o que pode levar a uma escolha equivocada a depender do nível de compreensão das perguntas.

O trabalho contribui para a literatura que estuda o uso de IA dentro do campo da contabilidade, em especial a aplicação de modelos de linguagem de processamento natural e LLMs. Muito tem se discutido sobre como esses modelos vão impactar a contabilidade, e considerando as limitações inerentes ao uso do Exame de Suficiência como uma *proxy* de problemas contábeis, entende-se que o uso desses modelos pode eventualmente ser integrado ao ambiente contábil, considerando também as suas limitações.

Modelos como o GPT podem alucinar (criando soluções simplesmente desconectadas da realidade) e variar suas respostas de maneira que não é possível compreender a razão exata dessa variabilidade. Os resultados podem contribuir com os desenvolvedores ao demonstrar um ponto de evolução ou de não evolução em conteúdos específicos, bem como evidenciar limitações da tecnologia ou dos avanços tecnológicos. Embora isso não tenha sido realizado na presente pesquisa, o ChatGPT permite que você elenque respostas erradas para a OpenAI e explique o que estava errado nessa questão. Esses *feedbacks* ao longo do tempo tem o potencial de aprimorar ainda mais modelos futuros.

A pesquisa apresenta como limitação a restrição dos problemas contábeis terem sido mensurados por meio de uma *proxy* (questões do Exame de Suficiência), a qual foi assumido como critério de aprovação dos indivíduos para o exercício profissional, ou seja, os requisitos dos profissionais. Sugere-se a expansão das análises para exames mais recentes e para outras atividades não parametrizadas para respostas em

múltipla escolha. Mais estudos sejam realizados que testem a eficácia dos modelos de PLN, inclusive, utilizando outras ferramentas como os modelos da Google (Gemini), da Meta (Meta AI) e outros LLMs que estão disponíveis.

Contribuição de autoria:

Freitas, M. M.: Conceituação, Metodologia, Software, Validação, Análise Formal, Pesquisa, Curadoria de dados, Escrita-rascunho original, Escrita, revisado e edição, Visualização, Supervisão, Administração do projeto. **Sallaberry, J. D.:** Conceituação, Metodologia, Validação, Análise Formal, Pesquisa, Escrita-rascunho original, Escrita, revisado e edição, Visualização, Supervisão, Administração do projeto. **Silva, T. B. J.:** Conceituação, Metodologia, Validação, Análise Formal, Pesquisa, Escrita-rascunho original, Escrita, revisado e edição, Visualização, Supervisão, Administração do projeto. **Rosa, F. S.:** Conceituação, Metodologia, Análise Formal, Pesquisa, Escrita-rascunho original, Escrita, revisado e edição, Visualização, Supervisão, Administração do projeto.

Marcelo Machado de Freitas (Freitas, M. M.)

Jonatas Dutra Sallaberry (Sallaberry, J. D.)

Thiago Bruno de Jesus Silva (Silva, T. B. J.)

Fabricia Silva da Rosa (Rosa, F. S.)

Declaração de conflito de interesse

Os autores declaram que durante o processo de pesquisa, não existiu nenhum tipo de interesse pessoal, profissional ou econômico que tenha podido influir no julgamento e/ou ações dos pesquisadores no momento de elaborar e publicar o artigo.

REFERÊNCIAS

- Albuquerque, F., & Gomes dos Santos, P. (2024). Can ChatGPT be a certified accountant? Assessing the responses of ChatGPT for the professional access exam in Portugal. *Administrative Sciences*, 14(7), 1-15. <https://doi.org/10.3390/admsci14070152>
- Atanasovski, A., Tocev, T., Dionisijev, I., Minovski, Z., & Jovevski, D. (2023). Evaluating the performance of ChatGPT in Accounting and Auditing Exams: An experimental study in North Macedonia. Em M Trpeska (Ed.), *Proceedings of the 4th International Conference Economic and Business Trends Shaping the Future* (pp. 40-50). Ss. Cyril and Methodius University. <http://doi.org/10.47063/EBTSF.2023.0003>

- 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. *Journal of AI*, 7(1), 52-62. <https://doi.org/10.61969/jai.1337500>
- Brasil. (2010, 11 de junho). *Lei nº 12.249, de 11 de junho de 2010. Altera a legislação tributária federal e dá outras providências*. Diário Oficial da União. https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2010/lei/l12249.htm
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Zieger, D. M., Wu, J., Winter, C.,... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Chui, M., Manyika, J., & Miremadi, M. (2016, 8 de julho). Where machines could replace humans—and where they can't (yet). *McKinsey Quarterly*. <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/where-machines-could-replace-humans-and-where-they-cant-yet>
- Conselho Federal de Contabilidade. (2022). *Exame de Suficiência: Panorama geral dos inscritos na 2ª edição de 2022*. CFC. <https://cfc.org.br/noticias/exame-de-suficiencia-panorama-geral-dos-inscritos-na-2a-edicao-de-2022>
- Conselho Federal de Contabilidade. (1999, 28 de julho). *Resolução CFC nº 853/1999. Institui o Exame de Suficiência como requisito para obtenção de Registro Profissional em Conselho Regional de Contabilidade*. Diário Oficial da União. <https://www.contabeis.com.br/legislacao/4629/resolucao-cfc-853-1999/>
- Conselho Federal de Contabilidade. (2023). *Exame de Suficiência*. CFC. <https://cfc.org.br/category/exame-de-suficiencia-anteriores/>
- Davenport, T. H., & Kirby, J. (2016). *Only humans need apply: Winners and losers in the age of smart machines*. Harper Business.
- Dong, M. M., Stratopoulos, T. C., & Wang, V. X. (2024). A scoping review of ChatGPT research in accounting and finance. *International Journal of Accounting Information Systems*, 55, 1-29. <https://doi.org/10.1016/j.accinf.2024.100715>
- Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers and Education: Artificial Intelligence*, 6, 1-13. <https://doi.org/10.1016/j.caeari.2023.100198>
- Eulerich, M., Sanatizadeh, A., Vakilzadeh, H., & Wood, D. A. (2024). Can artificial intelligence pass accounting certification exams? ChatGPT: CPA, CMA, CIA, and EA. *Review of Accounting Studies*, 29, 2318-2349. <https://doi.org/10.2139/ssrn.4452175>

Freitas, M. M., Sallaberry, J. D., Silva, T. B. & Rosa, F. S. (2024). Application of Chatgpt 4.0 for solving accounting problems. *GCG: Revista de Globalización, Competitividad y Gobernabilidad*, 18(2), 49-64.

Fu, X., Yang, L., Song, Y., & Yang, C. (2024). Exploration on the application of ChatGPT in corporate financial management. *Academic Journal of Management and Social Sciences*, 7(1), 65-73. <https://doi.org/10.54097/bb1xcw03>

Guerreiro, R., Amaral, J. V., Russo, P. T., & Mucci, D. M. (2023). Indústria 4.0: Características e potenciais impactos no ambiente interno das empresas. *Advances in Scientific and Applied Accounting*, 16(3), 92-104. <http://doi.org/10.14392/asaa.2023160305>

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 28(1), 75-105. <https://doi.org/10.2307/25148625>

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 1-9. <https://doi.org/10.1016/j.lindif.2023.102274>

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270), 1-17. <https://doi.org/10.1098/rsta.2023.0254>

Lima, C. E. S., Silva, A. C. B., Campelo, K. S., & Gonçalves, V. L. (2012, 26-29 de agosto). *Exame de Suficiência: Um estudo sobre a percepção de alunos concluintes de cursos de ciências contábeis do município de Caruaru-PE quanto a sua preparação face ao conhecimento adquirido na graduação* [Apresentação em conferência]. 19º Congresso Brasileiro de Contabilidade, Belém, Brasil.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187-1230. <https://doi.org/10.1111/1475-679X.12123>

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS One*, 13(3), 1-26. <https://doi.org/10.1371/journal.pone.0194889>

Mancebo, V.O.C., & Mucci, D. M. (2023, 26-28 de julho). *Impactos das tecnologias digitais nas atividades desempenhadas pela Controladoria* [Apresentação em conferência]. 23º USP International Conference on Accounting e 20º Congresso USP de Iniciação Científica em Contabilidade, São Paulo, Brasil.

Martínez, E. (2024). Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*, (2024), 1-24. <https://doi.org/10.1007/s10506-024-09396-9>

- Masuda, I., Nakagawa, K., & Hoshino, T. (2023, 6-9 de junho). *Can ChatGPT pass the JCPA exam?: Challenge for the short-answer method test on auditing* [Apresentação em conferência]. The 37th Annual Conference of the Japanese Society for Artificial Intelligence, Kumamoto, Japão.
- 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*, 9(2), 158-178. <https://doi.org/10.21680/2176-9036.2017v9n2ID10760>
- Nogueira, C. A. G., Nogueira, A. O., Almeida, S. F. D. C., Ribeiro, J. L., & Chaves, V. T. L. (2013, 25 de outubro). *Preparação para o Exame de Suficiência: Desafios para os docentes do curso de Ciências Contábeis* [Apresentação em conferência]. 5º Encontro de Práticas Docentes, Fortaleza, Brasil.
- Oliveira, J., & Khatib, A. (2023). Man or machine? An exploratory study of the performance of ChatGPT 3.5 in the CFC sufficiency exam. *Revista Capital Científico*, 22(1), 42-56. <https://doi.org/10.2139/ssrn.4560434>
- Oliveira Neto, J. D., & Kurotori, I. K. (2009). O retorno do Exame de Suficiência do CFC. *Revista de Contabilidade do Mestrado em Ciências Contábeis da UERJ*, 14(1), 62-79.
- OpenAI. (2024). *ChatGPT* [IA generativa]. OpenAI. <https://chat.openai.com>
- Pardos, Z. A., & Bhandari, S. (2024). Learning gain differences between ChatGPT and human tutor generated algebra hints. *PLoS One*, 19(5). 1-18. <https://doi.org/10.1371/journal.pone.0304013>
- Pursnani, V., Sermet, Y., Kurt, M., & Demir, I. (2023). Performance of ChatGPT on the US fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. *Computers and Education: Artificial Intelligence*, 5, 1-12. <https://doi.org/10.1016/j.caeari.2023.100183>
- Qadir, J. (2023). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. Em Institute of Electrical and Electronics Engineers (Ed.), *2023 IEEE Global Engineering Education Conference (EDUCON) Proceedings* (pp. 1-9). IEEE. <https://doi.org/10.36227/techrxiv.21789434.v1>
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI. chrome-extension://efaidnbmnnibpcajpcglclefindmkaj/https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- Sena, T. R., & Sallaberry, J. D. (2021). Contabilidade Geral e o desempenho dos estudantes no Exame de Suficiência: Uma pesquisa com IES baianas. *Revista de Contabilidade e Controladoria*, 13(1), 139-155. <https://doi.org/10.5380/rcc.v13i1.78581>

Terres, J. C., Simoni, P. R., Pereira, P., Timmermans, C., Lizote, S. A., & Lana, J. (2011, 7-9 de dezembro). *Exame de Suficiência da Profissão Contábil: Um estudo envolvendo o posicionamento de contabilistas e estudantes do curso de Ciências Contábeis a respeito de sua aplicabilidade* [Apresentação em conferência]. 11º Colóquio Internacional sobre Gestão Universitária na América do Sul, Florianópolis, Brasil.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Jacob, A. & Kim, Y. (2023). *Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks.* <https://arxiv.org/abs/2307.02477>

Data de recepção: 02/09/2025

Data de revisão: 02/09/2025

Data de aceitação: 09/09/2025

Contato: marcelo.freitas@ufsc.br