ARTIGO ORIGINAL

EFETIVIDADE DE CHATBOTS NO ENSINO EM CIÊNCIAS CONTÁBEIS: UMA ANÁLISE COMPARATIVA ENTRE *LLMs*

Carlos Roberto Souza Carmo[1]

Karla Aparecida Rabelo Fernandes[2]

Renata de Oliveira Souza Carmo[3]

RESUMO: Esta investigação teve como objetivo comparar o desempenho de diferentes *chatbots*, baseados em grandes modelos de linguagem (*LLM*), na resolução de questões de natureza contábil, abrangendo questionamentos de natureza binária, de múltipla escolha e questões abertas que exigiam raciocínio procedimental. Para tanto, foram analisadas as respostas fornecidas por sete modelos amplamente utilizados, incluindo o Claude, cujos resultados foram confrontados com aqueles obtidos por outros seis modelos já avaliados em pesquisa anterior. A investigação adotou abordagem qualitativa baseada em análises de caráter quantitativo, considerando tanto a precisão das respostas (acertos) quanto a capacidade dos modelos de interpretar enunciados, realizar cálculos encadeados e indicar corretamente a natureza de saldos contábeis. Os resultados sugerem que, embora o desempenho dos *chatbots* tenha sido semelhante nas questões binárias e nas questões de múltipla escolha, diferenças significativas emergiram nas questões abertas, nas quais o Claude apresentou melhor desempenho. A análise estatística realizada por meio do Teste Z para diferença entre proporções confirmou que essa superioridade foi significativa apenas nas questões abertas, indicando que o Claude pode se destacar em tarefas que exigem raciocínio contábil estruturado. Conclui-se que, apesar do potencial pedagógico dos *chatbots*, seu desempenho é heterogêneo e depende do tipo de tarefa, reforçando a necessidade de uso crítico e orientado dessas ferramentas no ensino das ciências contábeis.

PALAVRAS-CHAVE: Raciocínio procedimental; Modelos generativos; Educação digital.

[1] Doutor em Agronomia com ênfase em Energia na Agricultura pela Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP) (2020). Mestre em Ciências Contábeis pela Pontifícia Universidade Católica de São Paulo (PUC-SP) (2008). Possui especialização em Inteligência Artificial e Redes Neurais (2025), Ciência de Dados e *Big Data Analytics* (2024), *Data Mining* (2024) e Análise e Desenvolvimento de Sistemas em Python (2023). Professor da Faculdade de Ciências Contábeis da Universidade Federal de Uberlândia (FACIC-UFU). e-mail: carlosjj2004@hotmail.com. ORCID: https://orcid.org/0000-0002-3806-9228.

[2]. Graduação em Ciências Contábeis pela Universidade Federal de Uberlândia (FACIC-UFU). e-mail: karlaaprabelo@gmail.com. ORCID: https://orcid.org/0009-0006-5832-8326.

[3] Mestre em Educação pelo Programa de Pós-graduação em Educação da Universidade Federal de Uberlândia – PPGED-UFU (2018). Atua como professora de língua portuguesa e língua inglesa, suas literaturas e suas metodologias de ensino na Universidade de Uberaba e na Secretaria Municipal de Educação da Prefeitura de Uberaba-MG. e-mail: renatadeoliveira.carmo@gmail.com. ORCID: https://orcid.org/0000-0002-0997-0754.

ABSTRACT: This study aimed to compare the performance of different chatbots based on large language models in solving accounting-related questions, including binary items, multiple-choice questions, and open-ended problems requiring procedural reasoning. To this end, responses generated by seven widely used models—including Claude—were analyzed and compared with results from six other models previously evaluated in an earlier study. The investigation adopted a qualitative approach supported by quantitative analyses, considering both answer accuracy and the models' ability to interpret prompts, perform chained calculations, and correctly identify the nature of accounting balances. The findings indicate that, although chatbot performance was similar for binary and multiple-choice questions, substantial differences emerged in open-ended tasks, in which only Claude demonstrated superior performance. Statistical analysis using the Z-test for differences in proportions confirmed that this superiority was significant exclusively for open-ended questions, suggesting that Claude may excel in tasks requiring structured accounting reasoning. The study concludes that, despite the pedagogical potential of chatbots, their performance is heterogeneous and task-dependent, reinforcing the need for critical and guided use of these tools in accounting education.

KEYWORDS: Procedural reasoning; Generative models; Digital education.

1 INTRODUÇÃO

O cenário contemporâneo tem sido profundamente marcado pela incorporação acelerada de tecnologias capazes de transformara vida cotidiana de maneira contínua. Embora intensificado nos últimos anos, esse movimento remonta às transformações desencadeadas pela chamada *Indústria 4.0*, período em que emergiram tecnologias disruptivas destinadas a alterar processos tradicionais por meio da redefinição de como as pessoas executam suas atividades (Abreu *et al*., 2025; Shuhaiber; Kuhail; Salman, 2025). Nesse contexto de inovação permanente, a inteligência artificial (IA) vem assumindo papel central ao impulsionar o desenvolvimento de recursos cada vez mais sofisticados e amplamente acessíveis à sociedade (Adamopoulou; Moussiades, 2020).

O avanço de áreas como aprendizado de máquina (*machine learning*), visão computacional e processamento de linguagem natural tem ampliado significativamente o alcance da IA, que hoje permeia setores diversos, por exemplo, da saúde ao entretenimento, passando pela educação. Essa expansão tecnológica possibilita tanto a automação de tarefas rotineiras quanto o acesso rápido a grandes volumes de dados, favorecendo processos decisórios e otimizando atividades humanas (Carmo *et al*., 2025). No campo educacional, em particular, os *chatbots* baseados em grandes modelos de linguagem ou *Large Language Model* (*LLM*) têm ganhado destaque por sua capacidade de apoiar estudantes e docentes, oferecendo

esclarecimento de dúvidas, devolutivas imediatas sobre atividades e até mesmo experiências de aprendizagem personalizadas, ajustadas às necessidades individuais de cada usuário (Carmo *et al*., 2025).

Apesar de seu potencial, o uso dessas tecnologias não deve ocorrer de forma indiscriminada. A literatura evidencia que o desempenho dos *chatbots* pode variar substancialmente conforme o tipo de tarefa, apresentando resultados satisfatórios em algumas situações (Godke *et al*., 2024; Oliveira Jr.; Khatib, 2024), e ainda, limitações importantes em outras (Alves; Silva; Bonfim, 2024; Dong; Stratopoulos; Wang, 2024). Além disso, diferentes modelos podem demonstrar vantagens e fragilidades específicas, dependendo das finalidades para as quais são empregados (Kevian *et al*., 2024; Jiang; Gao; Karniadakis, 2025; Pascal, 2026).

Entre os sistemas mais utilizados e avaliados atualmente podem se destacar o Meta AI para WhatsApp (Meta AI, 2024), o Copilot integrado ao Office 365 (Microsoft, 2023b), o ChatGPT (OpenAI, 2023), o Copilot no navegador Edge (Microsoft, 2023a), o Gemini (Google AI, 2023), o DeepSeek-V3 (DeepSeek, 2024) e o Claude (Anthropic, 2025). Esses modelos têm sido submetidos a testes constantes e comparações frequentes, dada sua rápida evolução e ampla adoção. Assim, estudos comparativos envolvendo grandes modelos de linguagem (*LLM*) tornam-se fundamentais para compreender suas capacidades, limitações e adequações a diferentes contextos profissionais e educacionais.

Por exemplo, especificamente acerca do Claude (Anthropic, 2025), a literatura recente mostra que ele apresenta um conjunto de atributos que o posiciona de forma competitiva no ecossistema contemporâneo de *LLM*. Em áreas como controle clássico, estudos indicam que o modelo alcançou desempenho de última geração, destacando-se na resolução de problemas de graduação e superando resultados anteriormente obtidos por versões do GPT-4 em tarefas equivalentes (Kevian *et al*., 2024). Essa superioridade é atribuída, em parte, à capacidade dos modelos otimizados do Claude em integrar raciocínio matemático e fundamentos de engenharia, alcançando elevadas taxas de acurácia em avaliações especializadas (Kevian *et al*., 2024).

Outro eixo de vantagem refere-se ao desempenho do Claude em cenários que exigem reconhecimento da natureza estrutural de problemas científicos. Em experimentos comparativos, modelos otimizados para raciocínio, incluindo o Claude com pensamento estendido, têm demonstrado maior consistência na identificação de propriedades específicas dos problemas e na tomada de decisões alinhadas às exigências das tarefas em questão,

superando modelos de uso geral como DeepSeek V3 e ChatGPT 4, que por vezes ignoraram características essenciais, como rigidez ou instruções de implementação (Jiang; Gao; Karniadakis, 2025). Esse comportamento sugere que o Claude, quando configurado para raciocínio aprofundado, aproxima-se de estratégias humanas de resolução de problemas em computação científica e aprendizado de máquina (Jiang; Gao; Karniadakis, 2025).

Além disso, análises de natureza aplicada destacam que a Anthropic (2025) orienta o desenvolvimento do Claude para execução de tarefas complexas, e não para conversação casual. O modelo Claude Sonnet 4.5 é descrito como particularmente eficaz no seguimento literal de instruções complexas, evitando alucinações e demonstrando robustez em tarefas como análise de documentos extensos e escrita de código (Pascal, 2026). Recursos adicionais, como a funcionalidade de "uso do computador", ampliam seu potencial ao permitir a automação de tarefas operacionais em s*oftwares*, o que o torna especialmente adequado para profissionais como desenvolvedores, advogados e analistas de dados (Pascal, 2026). No campo dos agentes autônomos, o Claude é inclusive apontado como modelo de referência para sistemas que executam tarefas de forma independente (Pascal, 2026).

No entanto, apesar dessas vantagens, a literatura também identifica limitações importantes. Por exemplo, em estudos na área de endodontia, o Claude apresentou desempenho inferior ao ChatGPT-5 em termos de acurácia e completude das respostas, obtendo média de 3,44 (±1,19) em escala de cinco pontos, enquanto o concorrente alcançou 4,56 (±0,65), diferença estatisticamente significativa ($p = 0,002$) (Taşyürek; Adıgüzel; Ortaç, 2025). Esses resultados sugerem que, em domínios altamente especializados, o Claude pode apresentar lacunas relevantes (Taşyürek; Adıgüzel; Ortaç, 2025). No âmbito da integração tecnológica, o Claude também demonstra limitações. O Gemini apresenta vantagem competitiva ao se integrar de forma nativa ao ecossistema Google Workspace, facilitando fluxos de trabalho corporativos, enquanto o Claude opera de maneira mais independente, o que pode ser menos conveniente para organizações já inseridas em ambientes Google (Pascal, 2026).

Outra desvantagem observada refere-se à legibilidade dos textos produzidos. Em análises comparativas, o Gemini 2.5 Flash apresentou mediana de facilidade de leitura de Flesch (Flesch Reading Ease Score ou FRES, que pode ser entendido como um parâmetro utilizado para avaliar clareza, fluidez e complexidade textual produzida por *LLM*) de 31,1, considerada significativamente superior ao Claude Sonnet-4, que obteve apenas 8,3 (*p-valor* < 0,001), indicando que o Gemini produz textos mais acessíveis ao público geral (Taşyürek;

Adigüzel; Ortaç, 2025). Embora o Claude apresente legibilidade intermediária quando comparado ao Grok 4, que é um *LLM* que produz textos ainda mais complexos, sua posição permanece distante de modelos otimizados para acessibilidade textual (Taşyürek; Adigüzel; Ortaç, 2025).

Diante desse cenário e considerando que estudos recentes já analisaram o desempenho de diferentes *chatbots* em tarefas educacionais e informacionais na área de contabilidade, especificamente a pesquisa realizada por Carmo *et al*. (2025), o presente trabalho de pesquisa científica tem como objetivo geral comparar o desempenho do modelo Claude (Anthropic, 2025) com os resultados previamente identificados por Carmo *et al*. (2025) para o Meta AI para WhatsApp (Meta AI, 2024), o Copilot integrado ao Office 365 (Microsoft, 2023b), o ChatGPT (OpenAI, 2023), o Copilot no navegador Edge (Microsoft, 2023a), o Gemini (Google AI, 2023) e o DeepSeek V3 (DeepSeek, 2024).

Para isso, o Claude (Anthropic, 2025) foi submetido à mesma metodologia empregada por Carmo *et al*. (2025), permitindo uma análise comparativa rigorosa e sistemática entre os modelos, de modo a verificar em que medida seu desempenho se aproxima, supera ou diverge daquele observado nos demais sistemas avaliados anteriormente por Carmo *et al*. (2025). Nesse sentido, a presente investigação buscou responder ao seguinte questionamento direcionador: em que medida o desempenho do modelo Claude (Anthropic, 2025), quando submetido à mesma metodologia aplicada por Carmo *et al*. (2025), no contexto da educação contábil, se compara aos resultados obtidos pelos demais *chatbots* analisados, isto é, Meta AI para WhatsApp (Meta AI, 2024), Copilot no Office 365 (Microsoft, 2023b), ChatGPT (OpenAI, 2023), Copilot no navegador Edge (Microsoft, 2023a), Gemini (Google AI, 2023) e DeepSeek V3 (DeepSeek, 2024)?

A realização desta pesquisa se justifica, em primeiro lugar, sob uma perspectiva teórica, uma vez que a literatura sobre o uso de *chatbots* na educação contábil ainda se encontra em consolidação, especialmente no que diz respeito à compreensão comparativa das capacidades e limitações de diferentes modelos de linguagem em tarefas acadêmicas específicas. Do ponto de vista empírico, essa investigação se mostra necessária porque os resultados obtidos por Carmo *et al*. (2025) evidenciam variações significativas de desempenho entre os *chatbots* avaliados por eles, indicando que novos modelos, como o Claude (Anthropic, 2025), precisam ser examinados sob a mesma metodologia para que se possa verificar sua efetividade em relação aos demais *chatbots* já testados, tais como Meta AI para WhatsApp (Meta AI, 2024), Copilot no Office 365 (Microsoft, 2023b), ChatGPT

(OpenAI, 2023), Copilot no navegador Edge (Microsoft, 2023a), Gemini (Google AI, 2023) e DeepSeek V3 (DeepSeek, 2024). Por fim, a justificativa tecnológica para o desenvolvimento desse estudo decorre da rápida evolução dos grandes modelos de linguagem (*LLM*) e da necessidade de compreender como essas inovações podem ser aplicadas de forma segura, eficiente e alinhada às demandas da educação contábil, permitindo que docentes e estudantes se beneficiem de ferramentas mais precisas, confiáveis e adequadas às práticas pedagógicas contemporâneas.

2 REFERENCIAL TEÓRICO

Os grandes modelos de linguagem (*LLM*) utilizados em *chatbots* apresentam variações de desempenho que evidenciam a importância de compreender seus limites e potencialidades antes de aplicá-los no ensino superior em geral e, especialmente, no ensino das ciências contábeis. Embora alguns modelos, como o Claude, apresentem avanços relevantes em raciocínio matemático e científico (Kevian et al., 2024; Jiang; Gao; Karniadakis, 2025), estudos também apontam fragilidades de desempenho em domínios especializados e na clareza textual (Taşyürek; Adıgüzel; Ortaç, 2025). Assim, investigar essas diferenças de desempenho torna-se essencial para garantir que o uso dos *chatbots* baseados nesses *LLM* seja seguro, eficiente e alinhado às demandas formativas da educação.

O uso de *chatbots* baseados em inteligência artificial tem provocado transformações significativas no ambiente educacional, ampliando as possibilidades de ensino dinâmico, interativo e personalizado. Contudo, essa expansão tecnológica também tem sido acompanhada por desafios importantes, especialmente no que se refere à ocorrência de plágio, à imprecisão das respostas e à diminuição das interações humanas no processo de aprendizagem (Davar; Dewan; Zhang, 2025; Gökçearslan; Tosun; Erdemir, 2024). Esses aspectos revelam que, embora os *chatbots* ofereçam benefícios pedagógicos, sua adoção exige cautela e reflexão crítica por parte de instituições de ensino e educadores.

A preferência dos estudantes por recorrer aos *chatbots* para solucionar dúvidas, em detrimento ao diálogo com seus professores, tem sido apontada como um fator que reduz a interação interpessoal no ambiente acadêmico (Ferreira, 2025; Aguiar *et al*., 2024). Essa diminuição do contato humano compromete dimensões essenciais do desenvolvimento social e emocional, além de favorecer uma aprendizagem mecânica, pouco propícia ao debate e à construção coletiva do conhecimento (Ferreira, 2025; Aguiar *et al*., 2024). A dependência

crescente dessas ferramentas pode ainda limitar o pensamento crítico, uma vez que muitos estudantes passam a aceitar respostas prontas sem buscar compreender os fundamentos conceituais envolvidos (Memarian; Doleck, 2023).

As implicações éticas associadas ao uso de *chatbots* também merecem destaque, sobretudo no que diz respeito ao risco de plágio. A facilidade com que essas ferramentas tecnológicas geram textos acadêmicos pode incentivar a apropriação indevida de ideias, comprometendo a honestidade intelectual e a originalidade, princípios fundamentais no ensino superior (Fajt; Schiller, 2025; Elkhatat, 2023). Além disso, o plágio afeta a confiabilidade das avaliações, uma vez que trabalhos produzidos com auxílio indevido de assistentes virtuais podem não refletir o real nível de aprendizagem dos estudantes (Bringula, 2024; Dalalah; Dalalah, 2023).

Outro ponto crítico refere-se à confiabilidade das informações fornecidas pelos *chatbots*. Uma vez que esses modelos operam com base em padrões estatísticos e grandes bases de dados, podem gerar respostas incorretas ou descontextualizadas, especialmente no ambiente acadêmico, onde a regra tende a se pautar pela precisão conceitual (Baidoo-anu; Ansah, 2023; Memarian; Doleck, 2023). Esses erros, conhecidos como alucinações, consistem na produção de conteúdos aparentemente coerentes, mas sem respaldo factual adequado (Anh-Hoang; Tran; Nguyen, 2025; Zhang; Zhang, 2025). A ocorrência dessas falhas está relacionada tanto às limitações dos modelos quanto às ambiguidades presentes na comunicação humana, que podem levar a interpretações equivocadas por parte dos sistemas (Lemos, 2024; Barcellos; Albino, 2025).

Apesar da capacidade dos *chatbots* de gerar textos e imagens com rapidez e eficiência, é importante reconhecer que esses modelos tecnológicos não possuem compreensão real do conteúdo produzido. Por isso, a validação humana continua indispensável para assegurar a veracidade e a pertinência das informações apresentadas (Lima; Felippe, 2025; Barcellos; Albino, 2025; Currie, 2023). A confiança excessiva nessas ferramentas pode reduzir a motivação dos estudantes e levar a um envolvimento superficial com o aprendizado, prejudicando habilidades cognitivas essenciais, como análise crítica, interpretação de fatos e avaliação da credibilidade das fontes (Kooli, 2023; Zhai; Wibowo; Li, 2024; Akçapınar; Sidan, 2024).

Nesse cenário, o papel dos professores torna-se ainda mais estratégico. A presença dos *chatbots* na educação exige que os docentes adotem práticas pedagógicas que transcendam a simples transmissão de conteúdo, assumindo funções de mediação, orientação e *design* de

experiências de aprendizagem ativa (Berg; Plessis, 2023; Ling; Jan, 2025). Essa mudança implica estimular a autonomia intelectual dos estudantes, promovendo atividades que favoreçam a criatividade, o pensamento crítico e a pesquisa independente.

Dessa maneira, é fundamental que o processo educacional seja orientado para além da obtenção de respostas imediatas, e ainda, que busque valorizar a construção do conhecimento e o desenvolvimento de competências analíticas. Incentivar os estudantes a validar informações, questionar conteúdos gerados por IA e buscar fontes confiáveis contribui para uma formação mais sólida e alinhada às exigências contemporâneas do ensino superior (Fošner; Aver, 2025; Kooli, 2023). E, nesse sentido, os resultados dessa investigação podem ser especialmente úteis, não só pela comparação de desempenho entre um significativo número de *chatbots* amplamente utilizados atualmente, mas, também pela possibilidade de identificar aquele modelo que melhor se adapte à solução de questões de natureza procedimental tão comuns na contabilidade, ou seja problemas que envolvam regras técnicas, métodos próprios e estratégicas para alcançar os objetivos propostos (Godke *et al*., 2024). Isso por sua vez, pode fazer como que o uso de *chatbots* venha a ser compreendido, por estudantes de contabilidade, como um recurso complementar, e não como substituto da reflexão humana e da interação pedagógica.

3 METODOLOGIA DE PESQUISA

Ao avaliar comparativamente o desempenho de 6 *chatbots* diferentes aplicados na resolução de atividades acadêmicas próprias de um curso de graduação em ciências contábeis, Carmo *et al*. (2025) utilizaram atividades com questões de 3 tipologias/solicitações distintas, ou seja: atividade composta por dez questões binárias, nas quais eram propostas afirmações e solicitava-se uma resposta do tipo “verdadeiro” ou “falso”, doravante denominadas por “QVouF” - veja o Apêndice 1 da pesquisa de Carmo *et al*. (2025); atividade composta por dez questões de múltipla escolha, nas quais cada uma continha um enunciado expositivo e quatro possíveis alternativas, sendo somente uma correta, doravante denominadas por “QMult” - veja o Apêndice 2 da pesquisa de Carmo *et al*. (2025); e, atividade composta por dez questões abertas, para as quais foi fornecido um enunciado único para resolução geral, composto por 10 operações de naturezas diversas realizadas entre a matriz e uma filial da mesma empresa, e, ao final, foram apresentados dez questionamentos que demandaram respostas abertas, baseadas na apuração de saldos finais, a serem informados a partir da contabilização de todas

as operações descritas inicialmente, com apuração de resultados e elaboração de demonstrativos contábeis, doravante denominadas por “QAbert” - veja o Apêndice 3 da pesquisa de Carmo *et al*. (2025).

Uma vez disponíveis nos Apêndices do artigo produzido a partir da investigação realizada por Carmo *et al*. (2025), essas mesmas questões foram aplicadas ao Claude (Anthropic, 2025), de forma análoga ao procedimento realizado em relação ao Meta AI para WhatsApp (Meta AI, 2024), Copilot no Office 365 (Microsoft, 2023b), ChatGPT (OpenAI, 2023), Copilot no navegador Edge (Microsoft, 2023a), Gemini (Google AI, 2023) e DeepSeek V3 (DeepSeek, 2024).

Todavia, além do comparativo entre *chatbots* (“IA x IA”), Carmo *et al*. (2025) realizaram um comparativo “IA x humano”, no qual aplicaram aqueles 3 conjuntos de atividades (30 questões) para um grupo formado por 67 estudantes (humanos) de um curso de graduação em ciências contábeis. Após realizar análises qualitativas apoiadas em métodos quantitativos aplicados, Carmo *et al*. (2025) constataram que os *chatbots* apresentaram um desempenho semelhante ao dos alunos integrantes da sua amostra de pesquisa nas questões binárias e de múltipla escolha, e ainda, um desempenho ligeiramente inferior ao dos humanos nas questões abertas; porém, em todos os 3 grupos de questões os desempenhos foram considerados estatisticamente muito próximos.

Uma vez que o objetivo da presente investigação contempla uma análise comparativa exclusivamente entre os modelos de *chatbots* baseados em *LLM* (“IA x IA”), de modo a verificar em que medida o desempenho do Claude (Anthropic, 2025) se aproxima, supera ou diverge daquele desempenho observado para os 6 *chatbots* anteriormente avaliados por Carmo *et al*. (2025), não foram realizados procedimentos e análises aplicados a humanos. Ou seja, após a aplicação daqueles 3 conjuntos de atividades (30 questões) ao Claude (Anthropic, 2025), foi realizada uma análise comparativa de desempenho entre ele (Claude) o desemprenho observado e registrado por Carmo *et al*. (2025) para o Meta AI para WhatsApp (Meta AI, 2024), Copilot no Office 365 (Microsoft, 2023b), ChatGPT (OpenAI, 2023), Copilot no navegador Edge (Microsoft, 2023a), Gemini (Google AI, 2023) e DeepSeek V3 (DeepSeek, 2024).

Acerca do processo analítico propriamente dito, inicialmente, foi realizada uma análise qualitativa dos erros e acertos de cada um dos *chatbots* avaliados e, na sequência, procedeu-se à análise estatística para comparativo de desempenho geral (Claude x os demais *chatbots*). A comparação entre Claude (Anthropic, 2025) e o grupo formado por aqueles 6 *chatbots* foi

realizada por meio do teste Z para diferença entre duas proporções, adequado para situações em que se deseja avaliar se duas proporções independentes diferem de forma estatisticamente significativa.

As proporções de acertos foram calculadas conforme descrito na Equação 1, para o Claude (p_1) e para aquele grupo de 6 *chatbots* (p_2) avaliados incialmente por Carmo *et al*. (2025), ambos considerando os respectivos somatórios (∑) de acertos (x) e as respectivas quantidades totais de questões (n). A diferença entre proporções foi identificada por meio da Equação 2. A proporção combinada foi calculada conforme descrito pela Equação 3. O erro-padrão da diferença entre duas proporções amostrais foi calculado conforme descrito pela Equação 4. A Equação 5 descreve como foi calculada a estatística de teste Z. Finalmente, a Equação 6 identifica o nível de significância estatística por meio do *p-valor* bicaudal), sendo que, Φ representa a função de distribuição acumulada da normal padrão.

$$p_1 = \frac{\sum x_1}{n_1},\ p_2 = \frac{\sum x_2}{n_2} \quad (1)$$

$$p_1 - p_2 \quad (2)$$

$$p = \frac{\sum x_1 + \sum x_2}{n_1 + n_2} \quad (3)$$

$$SE = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad (4)$$

$$Z = \frac{p_1 - p_2}{SE} \quad (5)$$

$$p - valor = 2\,(1 - \Phi(|Z|)) \quad (6)$$

Considerando a natureza dos dados obtidos e os procedimentos analíticos empregados, este estudo configura-se como uma investigação científico-empírica de caráter qualitativo, sustentada por uma abordagem baseada em métodos quantitativos aplicados.

4 APRESENTAÇÃO, ANÁLISE E DISCUSSÃO DOS RESULTADOS

Ao iniciar o processo de análise de dados a partir dos resultados observados para as questões binárias, levou-se em conta que esse tipo de questionamento avalia essencialmente precisão factual, atenção ao enunciado e consistência lógica. Como são perguntas de baixa complexidade cognitiva, o

esperado é que modelos avançados apresentem desempenho elevado. Sendo que, isso se confirmou, conforme pode ser observado a partir dos dados resumidos no Quadro 1.

Os resultados indicam que o desempenho geral foi elevado entre os *chatbots* avaliados, o que é compatível com a baixa complexidade cognitiva exigida por esse tipo de tarefa (questões do tipo V ou F). Ainda assim, foram observadas diferenças que podem contribuir para a compreensão comparativa entre o Claude e os modelos previamente examinados por Carmo *et al*. (2025).

O modelo Claude apresentou desempenho satisfatório, com 10 acertos em 10 questões, igualando-se aos melhores resultados registrados no estudo anterior (Carmo *et al*., 2025), alcançados pelo Copilot no Office 365, Copilot no Edge, Gemini e DeepSeek V3. Esse desempenho evidencia a presença de estabilidade e ausência de oscilações interpretativas, reforçando a capacidade do Claude em lidar com afirmações factuais de forma precisa e consistente, assim como aconteceu com os modelos analisados inicialmente por Carmo *et al*. (2025), respeitado o respectivo contexto (formulação, tipo e conteúdo das questões propostas, entre outros fatores). Por outro lado, dois modelos avaliados por Carmo *et al*. (2025), o Meta AI e o ChatGPT, apresentaram desempenho ligeiramente inferior ao demais, com 9 acertos, ambos errando exclusivamente a questão QVouF-1. Essa coincidência no erro sugere que essa questão poderia possuir algum grau de ambiguidade ou exigia interpretação contextual mais refinada. Nesse ponto, o Claude demonstrou maior sensibilidade ao conteúdo da afirmação, distinguindo-se positivamente desses modelos.

Avançando para a avaliação qualitativa das questões de múltipla escolha, foi possível examinar a capacidade dos modelos em interpretar enunciados, discriminar alternativas plausíveis e selecionar a resposta correta, diante de informações potencialmente ambíguas ou concorrentes. Esse tipo de tarefa exige maior elaboração cognitiva do que as questões binárias, o que tende a ampliar as diferenças de desempenho entre os modelos.

Quadro 1 - Avaliação dos acertos e erros para as questões binárias (“IA *x* IA”)

Questão	Gabarito	Meta IA	Copilot no Office 365	ChaGPT	Copilot no Edge	Gemini	DeepSeek -V3	Claude
QVouF-1	V	*F*	*V*	*F*	*V*	*V*	*V*	*V*
QVouF-2	F	*F*	*F*	*F*	*F*	*F*	*F*	*F*
QVouF-3	V	*V*	*V*	*V*	*V*	*V*	*V*	*V*
QVouF-4	F	*F*	*F*	*F*	*F*	*F*	*F*	*F*
QVouF-5	F	*F*	*F*	*F*	*F*	*F*	*F*	*F*
QVouF-6	V	*V*	*V*	*V*	*V*	*V*	*V*	*V*
QVouF-7	V	*V*	*V*	*V*	*V*	*V*	*V*	*V*
QVouF-8	F	*F*	*F*	*F*	*F*	*F*	*F*	*F*
QVouF-9	F	*F*	*F*	*F*	*F*	*F*	*F*	*F*

QVouF-10	V	*V*	*V*	*V*	*V*	*V*	*V*	*V*
Total de acertos		9	10	9	10	10	10	10

Fonte: elaborado pelos autores a partir dos dados da pesquisa e em conjunto com os dados de Carmo *et al.* (2025).

Os resultados descritos no Quadro 2 evidenciam que o Claude obteve desempenho com 10 acertos nas 10 questões propostas, igualando-se ao melhor resultado observado entre os modelos avaliados anteriormente por Carmo *et al*. (2025), representado pelo Copilot no Office 365, que também acertou todas as questões.

Quadro 2 - Avaliação dos acertos e erros para as questões de múltipla escolha (“IA x IA”)

Questão	Gabarito	Meta IA	Copilot no Office 365	ChaGPT	Copilot no Edge	Gemini	DeepSeek-V3	Claude
QMult-1	d	*d*	*d*	*d*	*d*	*c*	*d*	*d*
QMult-2	a	*a*	*a*	*a*	*a*	*a*	*a*	*a*
QMult-3	c	*c*	*c*	*c*	*c*	*c*	*c*	*c*
QMult-4	b	*b*	*b*	*b*	*b*	*b*	*b*	*b*
QMult-5	a	*a*	*a*	*a*	*a*	*b*	*a*	*a*
QMult-6	d	*d*	*d*	*d*	*d*	*d*	*d*	*d*
QMult-7	c	*c*	*c*	*c*	*c*	*c*	*c*	*c*
QMult-8	c	*c*	*c*	*b*	*b*	*c*	*c*	*c*
QMult-9	c	*c*	*c*	*c*	*c*	*d*	*c*	*c*
QMult-10	c	*d*	*c*	*c*	*d*	*c*	*d*	*c*
Total de acertos		9	10	9	8	7	9	10

Fonte: elaborado pelos autores a partir dos dados da pesquisa e em conjunto com os dados de Carmo *et al.* (2025).

Entre os demais modelos avaliados por Carmo *et al*. (2025), observou-se uma variação significativa, segundo as informações resumidas no Quadro 2. Diferentemente do que ocorreu nas questões binárias, essa dispersão de resultados corretos pode ser um indício de que as questões de múltipla escolha expuseram fragilidades específicas de alguns modelos, respeitado o contexto da pesquisa em questão, especialmente em itens que envolviam alternativas parcialmente corretas ou que exigiam maior atenção a detalhes do enunciado. Todavia, qualquer inferência nesse sentido precisa ser avaliada com muito cuidado, especialmente, à luz da natureza e da forma dos questionamentos propostos, o que não foi alvo de análise neste estudo.

Ao realizar a análise qualitativa dos erros e acertos ocorridos no grupo das questões abertas, deve-se levar em conta que esse tipo de questão exige dos modelos a capacidade de interpretar um enunciado único, contabilizar diversas operações entre matriz e filial e, ao final, apurar os saldos das contas solicitadas, indicando sua natureza (devedora ou credora).

Trata-se do grupo mais complexo da avaliação, pois envolve raciocínio contábil sequencial, precisão numérica e coerência entre etapas.

Os resultados descritos no Quadro 3 demonstram um desempenho geral relativamente limitado entre os modelos avaliados por Carmo *et al*. (2025), respeitado o contexto proposto (formulação, tipo e conteúdo das questões propostas, entre outros fatores). Os erros observados incluem valores numericamente distantes do gabarito, ausência de resposta, inconsistências entre etapas de contabilização e indicação incorreta da natureza do saldo, o que pode ser um indício de dificuldades tanto na execução das operações quanto na interpretação do enunciado.

Em contraste, o Claude apresentou um desempenho significativamente melhor que os demais modelos avaliados anteriormente, com 8 acertos em 10 questões. O Claude parece manter coerência entre as etapas de contabilização dos fatos propostos, reproduzindo corretamente os saldos finais e indicar adequadamente sua natureza (devedora ou credora), em 80% dos casos (8 acertos em 10 questões). Mesmo nos itens em que houve divergência, suas respostas mantiveram proximidade com o gabarito, o que pode sugerir erros pontuais e não falhas estruturais de raciocínio; todavia, em contabilidade, erros dessa natureza podem ser desastrosos. Dessa maneira, justamente naquele tipo de questão que mais exige raciocínio contábil estruturado (nas questões abertas), o Claude se mostrou um modelo mais preciso e consistente, comparativamente ao desempenho dos *chatbots* avaliados anteriormente por Carmo *et al*. (2025).

Quadro 3 - Avaliação dos acertos e erros para as questões abertas (“IA x IA”)

Questão	Gabarito	Meta IA	Copilot no Office 365	ChaGPT	Copilot no Edge	Gemini	DeepSeek -V3	Claude
QAbert-1	162600 (d)	*162600 (d)*	*162600 (d)*	*162600 (d)*	*162600 (d)*	*162600 (d)*	*62600 (d)*	*162600 (d)*
QAbert-2	71000 (d)	*120000 (d)*	*71000 (d)*	*71000 (d)*	*102000 (d)*	*61000 (d)*	*51000 (d)*	*71000 (d)*
QAbert-3	23466,69 (c)	*0*	*23466,69 (c)*	*23466,69 (c)*	*31366,69 (c)*	*30716,69 (c)*	*23466,69 (c)*	*23466,69 (c)*
QAbert-4	0	*97940 (d)*	*0*	*0*	*0*	*Vlr. a ser calculado*	*21600 (c)*	*0*
QAbert-5	76780,52 (c)	*15000 (c)*	*57325,71 (c)*	*35292,99 (c)*	*89700 (c)*	*84800 (c)*	*15000 (c)*	*77330,52 (c)*
QAbert-6	46500 (c)	*16000 (c)*	*68000 (c)*	*54389,84 (c)*	*40900 (c)*	*49794,92 (c)*	*0*	*46500 (c)*
QAbert-7	123280,52 (c)	*31000 (c)*	*125325,7 (c)*	*89682,83 (c)*	*130600 (c)*	*134594,9 (c)*	*15000 (c)*	*123830,52 (c)*
QAbert-8	57900 (d)	*100000 (c)*	*100000 (d)*	*63100 (d)*	*50000 (c)*	*77900 (d)*	*50000 (d)*	*57900 (d)*
QAbert-9	57900 (c)	*100000 (d)*	*79000 (c)*	*63100 (c)*	*50000 (c)*	*77900 (c)*	*50000 (c)*	*57900 (c)*
QAbert-10	0	*754,14*	*650*	*650*	*650*	*Vlr. a ser*	*650*	*0*

		(d)	*(c)*	*(c)*	*(c)*	*calculado*	*(c)*	
Total de acertos		1	4	4	2	1	1	8

Fonte: elaborado pelos autores a partir dos dados da pesquisa e em conjunto com os dados de Carmo *et al.* (2025).

Após a análise qualitativa, que evidenciou diferenças relevantes entre o desempenho do Claude e dos demais modelos já avaliados por Carmo *et al.* (2025), procedeu-se à avaliação estatística dessas diferenças por meio do Teste Z para diferença entre duas proporções, considerando cada tipologia de questão separadamente. O objetivo desse procedimento foi determinar se as diferenças observadas qualitativamente se sustentavam estatisticamente.

Conforme descrito no Quadro 4, os resultados indicam que, nas questões binárias, embora o Claude tenha apresentado proporção de acertos ligeiramente superior (p_1 = 1,0000) à média dos demais modelos (p_2 = 0,9667), essa diferença não foi estatisticamente significativa (Z = 0,59; *p-valor* = 0,56). Esse achado confirma que o desempenho dos modelos é semelhante nesse tipo de tarefa, corroborando a interpretação qualitativa anterior. Em relação às questões de múltipla escolha, o Claude também obteve desempenho máximo (p_1 = 1,0000), que também foi superior a proporção média dos demais *chatbots* (p_2 = 0,8667). Contudo, essa diferença também não atingiu significância estatística (Z = 1,23; *p-valor* = 0,22). Assim, os resultados sugerem que, estatisticamente, o desempenho do Claude não se distancia de forma robusta dos demais modelos nesse tipo de questão.

Quadro 4 - Teste Z para diferença entre duas proporções

Chatbot / Parâmetros calculados	**Questões binárias**	**Questões de múltipla escolha**	**Questões abertas**
Claude	10	10	8
Meta IA	9	9	1
Copilot no Office 365	10	10	4
ChaGPT	9	9	4
Copilot no Edge	10	8	2
Gemini	10	7	1
DeepSeek-V3	10	9	1
p_1	1,0000	1,0000	0,8000
p_2	0,9667	0,8667	0,2167
p_1 - p_2	0,0333	0,1333	0,5833
p	0,9714	0,8857	0,3
SE	0,0569	0,1087	0,1565
Estatística Z	0,59	1,23	3,73
p-valor bicaudal (aprox.)	0,56	0,22	0,00

Legenda:
p_1 = proporção de acertos do Claude, por tipo de questão = ∑acertos / 10 questões;
p_2 = proporção de acertos dos demais chatbots (Meta IA+ Copilot no Office 365+ ChaGPT+ Copilot no Edge+ Gemini+ DeepSeek-V3), por tipo de questão = ∑acertos / 60 questões.

Fonte: elaborado pelos autores a partir dos dados da pesquisa e em conjunto com os dados de Carmo *et al*. (2025).

Por outro lado, nas questões abertas com um enunciado geral para resolução, observou-se um contraste mais expressivo entre o desempenho dos modelos analisados. O Claude apresentou proporção de acertos substancialmente superior (p_1 = 0,8000) à média dos demais *chatbots* (p_2 = 0,2167); diferença essa que se mostrou estatisticamente significativa (Z = 3,73; *p-valor* < 0,01). Esse resultado confirma, de forma quantitativa, aquela superioridade já identificada de forma qualitativa; isto é, os resultados sugerem que Claude é mais eficaz em tarefas que exigem raciocínio contábil estruturado, cálculos encadeados e interpretação detalhada do enunciado.

Os resultados desta investigação dialogam diretamente com as discussões teóricas sobre o desempenho dos grandes modelos de linguagem (*LLM*) e seus impactos no ensino superior. O desempenho do Claude nas questões abertas, especialmente naquelas que exigem raciocínio contábil estruturado, confirma observações de Kevian et al. (2024) e Jiang, Gao e Karniadakis (2025), que destacam avanços recentes de alguns *LLM* em tarefas matemáticas e científicas mais complexas. Ao mesmo tempo, o desempenho irregular conjunto formado pelos demais modelos reforça as limitações apontadas por Taşyürek, Adıgüzel e Ortaç (2025), que identificam possíveis fragilidades persistentes na atuação dos *chatbots* em domínios especializados. Porém, deve-se lembrar, mais uma vez, que inferências nesse sentido precisam ser avaliadas com muito cuidado, à luz da natureza e da forma dos questionamentos propostos, o que não foi alvo de análise neste estudo.

A despeito das ressalvas realizadas acerca da natureza e da forma dos questionamentos propostos, as possíveis dificuldades enfrentadas pelos modelos em manter coerência entre etapas de contabilização, cálculo e em interpretar adequadamente o enunciado também se alinham às preocupações de Baidoo-Anu e Ansah (2023) e Memarian e Doleck (2023) acerca da ocorrência de respostas imprecisas e alucinações. Nesse sentido, os achados também dialogam com as discussões sobre os riscos pedagógicos associados ao uso indiscriminado dessas ferramentas. Ou seja, a dependência excessiva dos estudantes em relação aos *chatbots*, apontada por Ferreira (2025) e Aguiar *et al*. (2024), torna-se ainda mais problemática quando se constata que muitos modelos não conseguem resolver adequadamente problemas procedimentais típicos da contabilidade. Ainda que de forma mais indireta, tais constatações também reforçam a preocupação de Memarian e Doleck (2023) de que o uso acrítico dessas

ferramentas pode limitar o desenvolvimento do pensamento crítico e favorecer a aceitação passiva de respostas incorretas.

A constatação de que apenas um modelo apresentou desempenho relativamente consistentemente reforça a importância da validação humana, conforme defendido por Lima e Felippe (2025) e Currie (2023). Mesmo quando um *chatbot* demonstra alta precisão, como os resultados dessa investigação sugerem em relação ao Claude, isso não elimina a necessidade de supervisão docente, já que a tecnologia não possui compreensão real do conteúdo produzido.

Por fim, os resultados desta pesquisa também dialogam com a perspectiva de Godke *et al*. (2024), que destacam a relevância de identificar ferramentas tecnológicas capazes de auxiliar na resolução de problemas procedimentais complexos. O desempenho do Claude sugere que determinados modelos podem, de fato, atuar como recursos complementares no ensino da contabilidade, desde que utilizados de forma crítica e orientada, preservando o papel do professor como mediador do processo de aprendizagem, conforme defendem Berg e Plessis (2023) e Ling e Jan (2025).

5 CONSIDERAÇÕES FINAIS

O resultado central desta pesquisa revela que o Claude apresentou desempenho relativo superior aos demais *chatbots* avaliados anteriormente. Sendo que, esse desempenhou mostrou-se mais evidente nas questões abertas, que exigiam raciocínio contábil estruturado e operações sequenciais.

Embora todos os modelos tenham alcançado resultados elevados nas questões binárias e desempenho relativamente próximo nas questões de múltipla escolha, os resultados observados sugerem que o Claude apresentou maior capacidade para interpretar corretamente o enunciado geral propostos para as questões abertas, realizar cálculos complexos e indicar adequadamente a natureza dos saldos contábeis, conforme demanda nesse tipos de questionamento. Esses achados podem sugerir que o Claude se destacaria como uma possível ferramenta apta a lidar com tarefas procedimentais típicas da contabilidade. Porém, para caminhar nesse sentido, ainda são necessários mais testes e avaliações.

A relevância desta pesquisa reside na comparação sistemática entre diferentes *chatbots* amplamente utilizados no contexto educacional, oferecendo evidências empíricas que podem orientar docentes, estudantes e instituições na escolha de ferramentas tecnológicas mais

adaptadas às demandas formativas. E, ao demonstrar que o desempenho dos modelos varia significativamente conforme o tipo de tarefa, o estudo contribui para uma compreensão mais crítica e fundamentada sobre o uso pedagógico da inteligência artificial no ensino das ciências contábeis. Além disso, os resultados reforçam a importância de considerar não apenas a fluidez textual dos *chatbots*, mas também sua precisão conceitual e capacidade de raciocínio, especialmente em atividades que envolvem cálculos e procedimentos técnicos.

Apesar das contribuições trazidas, esta pesquisa apresenta algumas limitações que devem ser reconhecidas. A primeira refere-se ao número reduzido de questões utilizadas na avaliação, e ainda, a respectiva delimitação temática (contabilidade comercial), o que, embora suficiente para identificar padrões relevantes, não abrange toda a diversidade de problemas contábeis encontrados na prática acadêmica e profissional. A segunda limitação diz respeito ao fato de que os modelos analisados estão em constante atualização, o que certamente alterará seu desempenho ao longo do tempo.

Para estudos futuros, entre outras possibilidades, sugere-se: ampliar o conjunto de questões, incluindo problemas de maior complexidade e de diferentes áreas da contabilidade; realizar análises longitudinais para verificar a evolução dos modelos; e, investigar como estudantes e professores utilizam esses *chatbots* em situações reais de aprendizagem, explorando impactos pedagógicos e éticos de forma mais aprofundada.

REFERÊNCIAS

ABREU, J. O. de S.; SANTIAGO, R. C.; NASCIMENTO FILHO, A. S.; CARDOSO, H. S. P.. Impacto da cultura organizacional na implementação de tecnologias disruptivas da indústria 4.0: uma revisão de literatura. **Revista de Gestão e Secretariado**, [*s.l.*], v. 16, n. 4, e-article4764, 2025. DOI: 10.7769/gesec.v16i4.4764. Disponível em: https://ojs.revistagesec.org.br/secretariado/article/view/4764. Acesso em: 06 dez. 2025.

ADAMOPOULOU, E.; MOUSSIADES, L.. Chatbots: History, technology, and applications. **Machine Learning with Applications**, [*s. l.*], v 2, e-article 100006, 15 December 2020. DOI: https://doi.org/10.1016/j.mlwa.2020.100006. Disponível em: https://www.sciencedirect.com/science/article/pii/S2666827020300062. Acesso em: 07 fev. 2025.

AGUIAR, M. do C. P. de; AZEVEDO, C. M. de S.; NASCIMENTO, J. S. do; CORRÊA, L. L.; BOTELHO, S. de O.. Educação a distância: vantagens, desvantagens e desafios da inserção da inteligência artificial. **Revista Ilustração**, Cruz Alta, v. 5, n. 5, p. 117-123, 2024. DOI: 10.46550/ilustracao.v5i5.336. Disponível em: https://journal.editorailustracao.com.br/index.php/ilustracao/article/view/336. Acesso em: 17 jan. 2026.

AKÇAPINAR, G.; SIDAN, E.. AI chatbots in programming education: guiding success or encouraging plagiarism. **Discover Artificial Intelligence**, [*s.l.*], v. 4, e-article 87, 2024. DOI: https://doi.org/10.1007/s44163-024-00203-7. Disponível em: https://link.springer.com/article/10.1007/s44163-024-00203-7. Acesso em: 17 jan. 2026.

ALVES, M. A.; SILVA, C. A. T.; BONFIM, M. P.. ChatGPT e desonestidade acadêmica: percepção dos estudantes de contabilidade sobre o seu uso. **Revista GUAL**, Florianópolis, v. 17, n. 3, e-article 99115, 2025. DOI: https://doi.org/10.5007/1983-4535.2024.e99115. Disponível em: https://periodicos.ufsc.br/index.php/gual/article/view/99115. Acesso em: 21 jan. 2026.

ANH-HOANG, D.; TRAN, V.; NGUYEN, Le-M.. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. **Frontiers Artificial Intelligence**, [*s.l.*], v. 8, e-article 1622292, 2025. DOI: 10.3389/frai.2025.1622292. Disponível em: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1622292. Acesso em: 16 jan. 2026.

ANTHROPIC. **Claude**. Versão Sonnet 4.5. San Francisco: Anthropic, 2025. Assistente de inteligência artificial baseado em modelo de linguagem. Disponível em: https://claude.ai. Acesso em: 30 jan. 2026.

BAIDOO-ANU, D.; ANSAH, L. O.. Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. **Journal of AI**, [*s.l.*], v. 7, n. 1, p. 52-62, 2023. DOI: https://doi.org/10.61969/jai.1337500. Disponível em: https://dergipark.org.tr/en/pub/jai/article/1337500. Acesso em: 07 jan. 2026.

BARCELLOS, L. I.; ALBINO, J. P.. Desinformação e inteligência artificial: Impacto das alucinações na utilização do ChatGPT para a área acadêmica. **ARACÊ**, [*s.l.*], v. 7, n. 12, e-article 11100, 2025. DOI: https://doi.org/10.56238/arev7n12-164. Disponível em: https://periodicos.newsciencepubl.com/arace/article/view/11100. Acesso em: 14 jan. 2026.

BERG, G. V. D.; PLESSIS, E. D.. ChatGPT and generative AI: possibilities for its contribution to lesson planning, critical thinking and openness in teacher education. **Education Sciences**, [*s.l.*], v. 13, n. 10, e-article 998, 2023. DOI: https://doi.org/10.3390/educsci13100998. Disponível em: https://www.mdpi.com/2227-7102/13/10/998. Acesso em: 13 jan. 2026.

BRINGULA, R.. ChatGPT in a programming course: Benefits and limitations. **Frontiers in Education**, [*s.l.*], v. 9, e-article 1248705, 2024. DOI: https://doi.org/10.3389/feduc.2024.1248705. Disponível em: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2024.1248705/full. Acesso em: 08 jan. 2026.

CARMO, C. R. S.; CARMO, R. de O. S.; ÁVILA, M. F. P. de; ÁVILA, L. A. C. de. Inteligência artificial (IA) e ensino superior: análise de desempenhos “IA versus IA” e “IA versus humano. **Cadernos da Fucamp**, Monte Carmelo, v. 40, p. 84-113, 2025. Disponível em: https://revistas.fucamp.edu.br/index.php/cadernos/article/view/3778. Acesso em: 17 dez. 2025.

CURRIE, G. M.. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? **Seminars in Nuclear Medicine**, [*s.l.*], v. 53, n. 5, p. 719-730, 2023. DOI:

https://doi.org/10.1053/j.semnuclmed.2023.04.008. Disponível em: https://www.sciencedirect.com/science/article/pii/S0001299823000363. Acesso em: 14 jan. 2026.

DALALAH, D.; DALALAH, O. M. A.. The false positives and false negatives of generative AI detection tools in education and academic research: the case of ChatGPT. **The International Journal of Management Education**, [*s.l.*], v. 21, n. 2, e-article 100822, 2023. DOI: https://doi.org/10.1016/j.ijme.2023.100822. Disponível em: https://www.sciencedirect.com/science/article/pii/S1472811723000605. Acesso em: 08 jan. 2026.

DAVAR, N. F.; DEWAN, M. A. A.; ZHANG, X.. AI chatbots in education: challenges and opportunities. **Information**, [*s.l.*], v. 16, n. 3, p. 235-260, 2025. DOI: http://dx.doi.org/10.3390/info16030235. Disponível em: https://www.mdpi.com/2078-2489/16/3/235. Acesso em: 12 dez. 2025.

DEEPSEEK. **DeepSeek-V3**: modelo de linguagem de inteligência artificial. [*S. l.*]: DeepSeek, Dec. 2024. Exercícios sobre Contabilidade e Direito Empresaria. Data da consulta: 04 fev. 2025-15h15.Disponível em: https://www.deepseek.com. Acesso em: 05 fev. 2025-08h40.

DONG, M. M.; STRATOPOULOS, T. C.; WANG , V. X.. A scoping review of ChatGPT research in accounting and finance. **International Journal of Accounting Information Systems**, [*s.l.*], v. 55, e-article 100715, 2024. DOI: https://doi.org/10.1016/j.accinf.2024.100715. Disponível em: https://www.sciencedirect.com/science/article/pii/S1467089524000484. Acesso em: 21 jan. 2026.

ELKHATAT, A. M.. Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities. **International Journal for Educational Integrity**, [*s.l.*], v. 19, n. 15, 2023. DOI: https://doi.org/10.1007/s40979-023-00137-0. Disponível em: https://link.springer.com/article/10.1007/s40979-023-00137-0. Acesso em: 13 jan. 2026

FAJT, B.; SCHILLER, E.. ChatGPT in academia: University students' attitudes towards the use of ChatGPT and plagiarism. **Journal of Academic Ethics**, [*s.l.*], v. 23, p. 1363-1382, 2025. DOI: https://doi.org/10.1007/s10805-025-09603-5. Disponível em: https://link.springer.com/article/10.1007/s10805-025-09603-5. Acesso em: 13 jan. 2026.

FERREIRA, R. P.. A inteligência artificial na educação: entre os benefícios e os riscos. **Revista Tópicos**, [*s.l.*], v. 3, n. 24, e-article 16990063, 2025. DOI: 10.5281/zenodo.16990063. Disponível em: https://zenodo.org/records/16990063. Acesso em: 11 jan. 2026.

FOŠNER, A.; AVER, B.. AI chatbots in higher education: students' beliefs and concerns. **Sustainable Futures**, [*s.l.*], v. 9, e-article 100734, 2025. DOI: http://dx.doi.org/10.1016/j.sftr.2025.100734. Disponível em: https://www.sciencedirect.com/science/article/pii/S2666188825003004. Acesso em: 12 dez. 2025.

GODKE, R. de F. G.; SILVA, O. L. da; COLAUTO, R. D.; CUNHA, J. V. A. da; DURSO, S. de O.. Sucesso ou fracasso? Desempenho do chatgpt nas habilidades conceituais, procedimentais e atitudinais do exame de suficiência do CFC. **Revista Catarinense da**

Ciência Contábil, Florianópolis, SC, v. 23, e-article 3525, 2024. DOI: https://doi.org/10.16930/2237-766220243525. Disponível em: https://revista.crcsc.org.br/CRCSC/article/view/3525. Acesso em: 21 jan. 2026.

GOOGLE AI. **Google Gemini (2.0 Flash)**. [*S. l.*]: Google AI, Dec. 2023. Data da consulta: 04 fev. 2025-14h44. Disponível em: https://gemini.google.com/app/f8a86e025d888c7d?hl=pt-PT. Acesso em: 05 fev. 2025-08h40.

GÖKÇEARSLAN, S.; TOSUN, C.; ERDEMIR, Z. G.. Benefits, challenges, and methods of artificial intelligence (AI) chatbots in education: a systematic literature review. **International Journal of Technology in Education**, [*s.l.*], v. 7, n. 1, p. 19-39, 2024. DOI: https://doi.org/10.46328/ijte.600. Disponível em: https://eric.ed.gov/?id=EJ1415037. Acesso em: 10 jan. 2026

JIANG, Q.; GAO, Z. ; KARNIADAKIS, G. E.. DeepSeek vs. ChatGPT vs. Claude: a comparative study for scientific computing and scientific machine learning tasks. **Theoretical and Applied Mechanics Letters**, [*s. l.*], v. 15, issue 3, e-article 100583, 2025. DOI: https://doi.org/10.1016/j.taml.2025.100583. Disponível em: https://www.sciencedirect.com/science/article/pii/S2095034925000157. Acesso em: 31 jan. 2026

KEVIAN, D.; SYED, U.; GUO, X.; HAVENS, A.; DULLERUD, G.; SEILER, P.; QIN, L.; HU. B..Capabilities of large language models in control engineering: a benchmark study on GPT-4, Claude 3 Opus, and Gemini 1.0 Ultra. **arXiv [math.OC],** [*s. l.*], e-artilcle 2404.03647v1 Apr. 2024. DOI: https://doi.org/10.48550/arXiv.2404.03647. Disponível em: https://arxiv.org/abs/2404.03647. Acesso em: 31 jan. 2026.

KOOLI, C.. Chatbots in education and research: a critical examination of ethical implications and solutions. **Sustainability**, [*s.l.*], v. 15, n. 7, e-article 5614, 2023. DOI: https://doi.org/10.3390/su15075614. Disponível em: https://www.mdpi.com/2071-1050/15/7/5614. Acesso em: 16 dez 2025.

LEMOS, A. L. M.. Erros, falhas e perturbações digitais em alucinações das IA generativas: Tipologia, premissas e epistemologia da comunicação. **MATRIZes**, São Paulo, Brasil, v. 18, n. 1, p. 75–91, 2024. DOI: 10.11606/issn.1982-8160.v18i1p75-91. Disponível em: https://revistas.usp.br/matrizes/article/view/210892. Acesso em: 03 jan. 2026.

LIMA, J. C. R. de; FELIPPE, M. L.. Integridade acadêmica na era do ChatGPT, desafios éticos e as novas fronteiras da Inovação. **Revista Brasileira da Educação Profissional e Tecnológica**, [*s.l.*], v. 3, n. 25, e-article 17803, 2025. DOI: https://doi.org/10.15628/rbept.2025.17803. Disponível em: https://www2.ifrn.edu.br/ojs/index.php/RBEPT/article/view/17803. Acesso em: 13 jan. 2026.

LING, Y.; JAN, J. M.. Voices from the flip: teacher perspectives on integrating AI chatbots in flipped english classrooms. **Education Sciences**, [*s.l.*], v. 15, n. 9, e-article 1219, 2025. DOI: https://doi.org/10.3390/educsci15091219. Disponível em: https://www.mdpi.com/2227-7102/15/9/1219. Acesso em: 16 jan. 2026.

MEMARIAN, B.; DOLECK, T.. ChatGPT in education: methods, potentials, and limitations. **Computers in Human Behavior: Artificial Humans**, [*s.l.*], v. 1, n. 2, e-article 103191, 2023.

DOI: https://doi.org/10.1016/j.chbah.2023.100022. Disponível em: https://www.sciencedirect.com/science/article/pii/S2949882123000221. Acesso em: 07 jan. 2026.

META AI. **Llama 3.2: modelo de linguagem desenvolvido pela Meta AI**. [*S. l.*]: Meta AI, Oct. 2024. Respostas às perguntas sobre Sociedade Limitada. Data da consulta: 04 fev. 2025-14h25. Consulta feita via WhatsApp. Acesso em: 05 fev. 2025-08h40.

MICROSOFT. **Microsoft Copilot no Edge**. [*S. l.*]: Microsoft, Feb. 2023a. Data da consulta: 04 fev. 2025-14h40. Disponível em: https://www.microsoft.com/en-us/edge/copilot?form=MA13RM. Acesso em: 05 fev. 2025-08h40.

MICROSOFT. **Microsoft Copilot no Office 365**. [*S. l.*]: Microsoft, Mar. 2023b. Data da consulta: 04 fev. 2025-14h26. Disponível em: https://www.microsoft.com/en-us/ai/copilot. Acesso em: 05 fev. 2025-08h40.

OLIVEIRA Jr., J. C. R. de; KHATIB, A. S. El. Homem ou máquina? um estudo exploratório do desempenho do ChatGP 3.5 no exame de suficiência do CFC. **Revista Capital Científico – Eletrônica (RCCe)**, [*s.l.*], v. 22, n.1, p. 42-56, 2024. DOI: https://doi.org/10.5935/2177-4153.20240003. Disponível em: https://revistas.unicentro.br/index.php/capitalcientifico/article/view/7609. Acesso em: 24 jan. 2026.

OPENAI. **ChatGPT (modelo GPT-4)**. [*S. l.*]: OpenAI , Mar. 2023. Data da consulta: 04 fev. 2025-14h34. Disponível em: https://openai.com. Acesso em: 05 fev. 2025-08h40.

PASCAL, Frederiek. Gemini 3 Pro vs ChatGPT 5.2 vs Claude Sonnet 4.5 vs Perplexity: the definitive choice for 2026. **CLICKFOREST**, [*s. l.*], AI 2026. Disponível em: https://www.clickforest.com/en/blog/gemini-3-pro-vs-chatgpt-vs-claude-vs-perplexity. Acesso em: 30 jan. 2026.

SHUHAIBER, A.; KUHAIL, M. A.; SALMAN, S.. ChatGPT in higher education - a student's perspective. **Computers in Human Behavior Reports**, [*s.l.*], v. 17, e-article 100565, 2025. DOI: https://doi.org/10.1016/j.chbr.2024.100565. Disponível em: https://www.sciencedirect.com/science/article/pii/S2451958824001982. Acesso em: 16 dez. 2025.

TAŞYÜREK, Makbule; ADIGÜZEL, Özkan; ORTAÇ, Hatice. Comparative evaluation of responses from ChatGPT-5, Gemini 2.5 Flash, Grok 4, and Claude Sonnet-4 chatbots to questions about endodontic iatrogenic events. **Healthcare**, [*s. l.*] v. 13, n. 20, p. 2615, 2025. DOI: 10.3390/healthcare13202615. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC12562575/. Acesso em: 30 jan. 2026.

ZHAI, C.; WIBOWO, S.; LI, L. D.. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. **Smart Learning Environments**, [*s.l.*], v. 11, e-article 28, 2024. DOI: https://doi.org/10.1186/s40561-024-00316-7. Disponível em: https://link.springer.com/article/10.1186/s40561-024-00316-7. Acesso em: 16 jan. 2026.

ZHANG, W.; ZHANG, J.. Hallucination mitigation for retrieval-augmented large language models: a review. **Mathematics**, [*s.l.*], v. 13, n. 5, e-article 856, 2025. DOI:

https://doi.org/10.3390/math13050856. Disponível em: https://www.mdpi.com/2227-7390/13/5/856. Acesso em: 16 jan. 2026.