

RELATÓRIOS GERENCIAIS E EXAME DE SUFICIÊNCIA: UMA ANÁLISE DOS CONTEÚDOS DAS DIRETRIZES CURRICULARES DE CIÊNCIAS CONTÁBEIS FRENTE ÀS EXIGÊNCIAS PROFISSIONAIS

Lucas Gabriel Slompo¹
Giovani Antônio Bordini²

RESUMO

O uso crescente da Inteligência Artificial Generativa levanta questões sobre a capacidade técnica desses sistemas em exames de qualificação profissional. Este estudo analisa o desempenho de dois Modelos de Linguagem de Grande Porte (LLMs), Gemini 3.1 Flash-Lite e GPT-5.4 Nano, na resolução e classificação de 434 questões válidas do Exame de Suficiência do Conselho Federal de Contabilidade (CFC) entre 2022 e 2026. Com base na Teoria do Capital Humano e na Economia da Sinalização, a pesquisa utilizou *Retrieval-Augmented Generation* (RAG) para fornecer suporte teórico aos modelos, simulando o acesso a normas e livros-texto. Os resultados indicam desempenho excepcional do Gemini, que alcançou precisão de 89,40% em zero-shot e 89,63% com RAG. Esses números superam significativamente as taxas de aprovação humana, que variaram entre 12,1% e 48,2% no período. O GPT obteve 74,65% de precisão, mostrando capacidade de aprovação, mas evidenciando diferenças entre as ferramentas. O RAG gerou “alucinação induzida” no GPT, reduzindo seu desempenho para 73,04%, o que confirma teorias sobre o excesso de contexto. A concordância inter-modelos (Kappa de Cohen) foi alta na classificação por área (Kappa = 0,77), confirmando a concentração em Contabilidade Geral. No entanto, foi baixa na autoavaliação de dificuldade e na aderência às Diretrizes Curriculares Nacionais. Conclui-se que os LLMs possuem capacidade técnica para atuar como tutores especialistas, exigindo reflexão dos órgãos reguladores sobre a avaliação dos futuros profissionais contábeis diante da automação.

Palavras-chave: Inteligência Artificial. Exame de Suficiência. Modelos de Linguagem. Educação Contábil. RAG.

- 1 Aluno do 6º período do curso de Economia da FAE Centro Universitário. Bolsista do Programa de Apoio à Iniciação Científica (PAIC 2024-2025). *E-mail:* lucas.slompo@mail.fae.edu
- 2 Orientador da Pesquisa. Bacharel em Ciências Contábeis e Doutor em Engenharia de Produção. Professor da FAE Centro Universitário. *E-mail:* giovani.bordini@fae.edu

INTRODUÇÃO

A contabilidade contemporânea vai muito além do simples registro de operações e do cumprimento de regras tributárias. Hoje, ela tem um papel central e estratégico na produção de informações úteis para a gestão, o controle de riscos e a tomada de decisões nas empresas (Padoveze, 2021). No Brasil, o Exame de Suficiência, organizado pelo Conselho Federal de Contabilidade (CFC), é o marco obrigatório para a obtenção do registro profissional. Ele funciona como o principal mecanismo legal para garantir a qualidade técnica dos contadores que entram no mercado de trabalho (Brasil, 2010).

Apesar de sua importância, as taxas de aprovação no exame apresentam uma variação muito grande e, frequentemente, níveis preocupantes. Em edições recentes, como no segundo semestre de 2023 e de 2024, a aprovação de bacharéis em Ciências Contábeis chegou perto de apenas 12% (Conselho Federal de Contabilidade [CFC], 2024). Essa realidade levanta debates importantes sobre a eficiência dos métodos de ensino e se a formação acadêmica está alinhada com as Diretrizes Curriculares Nacionais (DCNs), recentemente atualizadas pela Resolução CNE/CES nº 1/2024 (Brasil, 2024).

Paralelo aos desafios na educação contábil, o mercado acompanha o rápido avanço da Inteligência Artificial Generativa. Modelos de Linguagem de Grande Porte (conhecidos pela sigla em inglês LLMs) mostram uma capacidade cada vez maior de interpretar normas, analisar relatórios e resolver problemas complexos (Eulerich; Masli; Wood, 2023). Testes recentes em exames profissionais difíceis, como provas da área de direito, de medicina e exames contábeis internacionais (como CPA e CMA nos Estados Unidos), demonstraram que as máquinas já conseguem superar a média de acertos dos candidatos humanos (Eulerich; Masli; Wood, 2023). No Brasil, no entanto, ainda faltam estudos abrangentes que testem essas tecnologias especificamente no exame do CFC, especialmente considerando a complexidade dos Pronunciamentos Técnicos (CPCs) e das Normas Brasileiras de Contabilidade (NBCs).

Diante desse contexto, o presente trabalho levanta a seguinte questão de pesquisa: os modelos de inteligência artificial de grande porte, quando configurados como especialistas em contabilidade, são capazes de atingir taxas de acerto no Exame de Suficiência do CFC comparáveis ou superiores às dos candidatos humanos? Além disso, de que forma o uso de bases locais de consulta, utilizando a técnica de Retrieval-Augmented Generation (RAG), afeta esse desempenho e a capacidade dos modelos de avaliar e classificar as questões?

O objetivo principal desta pesquisa é analisar o desempenho de modelos de inteligência artificial generativa de ponta (especificamente Gemini do Google e GPT da OpenAI) na resolução e classificação de questões do Exame de Suficiência do CFC,

aplicadas entre 2022 e 2026. A partir desses dados, busca-se comparar os resultados técnicos da IA com as taxas de aprovação dos estudantes, testando também a consistência dos modelos quando atuam como avaliadores pedagógicos.

Para apoiar essa análise, os objetivos específicos são: a) submeter 434 questões válidas de 9 edições consecutivas aos modelos selecionados; b) avaliar a precisão de resolução com e sem acesso a materiais de consulta; c) solicitar que a inteligência artificial classifique as questões por nível de dificuldade, aderência às DCNs e área de conhecimento; d) utilizar a métrica estatística Kappa de Cohen (Cohen, 1960) para verificar se diferentes modelos concordam em suas avaliações.

A justificativa deste estudo baseia-se na necessidade de entender os limites e as oportunidades da inteligência artificial na educação contábil. Na visão da Teoria do Capital Humano e da Economia da Sinalização (Becker, 1964; Spence, 1973), o diploma e o registro profissional servem para sinalizar ao mercado de trabalho que o estudante possui a competência necessária. Se a tecnologia consegue resolver com facilidade provas que são consideradas complexas para os alunos, surge uma oportunidade de usar a IA como uma ferramenta de ensino. Ao mesmo tempo, questiona-se se o exame do CFC avalia competências que podem ser facilmente automatizadas, o que iria contra as novas diretrizes do MEC, focadas em formar profissionais com capacidade de julgamento crítico e tomada de decisão (Brasil, 2024).

Além desta introdução, o artigo está dividido em mais quatro seções. A seção 2 apresenta a revisão da literatura teórica e técnica. A seção 3 detalha a metodologia de pesquisa e os modelos computacionais utilizados. A seção 4 apresenta e discute os resultados do experimento. Por fim, a seção 5 resume as considerações finais e as limitações do estudo.

1 REVISÃO DA LITERATURA

1.1 O EXAME DE SUFICIÊNCIA E AS DIRETRIZES CURRICULARES NACIONAIS

O Exame de Suficiência do CFC foi instituído de forma definitiva pela Lei nº 12.249/2010. Ele funciona como um filtro para o mercado de trabalho, garantindo que o profissional recém-formado possua os conhecimentos técnicos mínimos necessários (Brasil, 2010).

Para analisar a função desse exame sob um ponto de vista econômico, a Teoria do Capital Humano de Gary Becker (1964) e a Teoria da Sinalização de Michael Spence

(1973) são essenciais. Segundo Spence (1973), o mercado de trabalho possui uma assimetria de informações. Em outras palavras, os empregadores não sabem ao certo o nível real de conhecimento de um candidato apenas olhando para o seu diploma. Para resolver isso, o candidato precisa apresentar um “sinal”. Como a qualidade das instituições de ensino varia muito, o Exame de Suficiência funciona como o grande sinalizador de qualidade: a aprovação mostra ao mercado que o profissional tem competência comprovada.

No entanto, o conteúdo cobrado na prova precisa estar alinhado com o que o Ministério da Educação (MEC) define como o padrão de formação ideal. Esse padrão é estabelecido pelas Diretrizes Curriculares Nacionais (DCNs). A Resolução de 2004 moldou o ensino contábil por muito tempo, exigindo conhecimentos de contabilidade básica, financeira e gerencial (Brasil, 2004). O problema é que, historicamente, muitos pesquisadores, como Peleias *et al.* (2007) e Padoveze (2021), alertam para um ensino muito focado apenas na memorização de regras e na mecânica de débitos e créditos, deixando a análise gerencial em segundo plano.

Buscando modernizar a profissão, a nova Resolução CNE/CES nº 1/2024 trouxe mudanças importantes. Ela foca no uso de tecnologia, na análise de dados e nas habilidades comportamentais e éticas necessárias para tomar decisões complexas (Brasil, 2024). O grande debate que surge é se as provas atuais do Conselho Federal de Contabilidade (CFC), agora organizadas pela Fundação Getúlio Vargas (FGV), conseguem avaliar essas novas habilidades ou se continuam medindo apenas a memorização de normas (CFC, 2024).

1.2 INTELIGÊNCIA ARTIFICIAL GENERATIVA E LLMS

Enquanto as universidades discutem como atualizar seus currículos, o mundo passou por uma transformação tecnológica com os Modelos de Linguagem de Grande Porte (LLMs). Estes sistemas são baseados em redes neurais artificiais, especificamente na arquitetura Transformer, criada pelo Google em 2017.

Modelos como o GPT (da OpenAI) e o Gemini (do Google) funcionam a partir de uma lógica estatística: eles foram treinados com um volume gigantesco de textos e aprenderam a prever qual será a próxima palavra mais provável em uma frase. Como essas redes neurais possuem bilhões de parâmetros (as “conexões” do cérebro artificial), elas conseguem entender contextos complexos. É por isso que uma IA consegue interpretar leis societárias, traduzir textos e escrever relatórios sem ter sido programada passo a passo para fazer isso (Topsakal; Akinci, 2023).

Apesar de muito avançada, essa tecnologia possui uma falha conhecida no meio acadêmico como “alucinação”. Como os LLMs apenas calculam probabilidades matemáticas entre palavras, quando perguntados sobre um assunto muito específico que eles não aprenderam direito no treinamento, eles podem inventar respostas e fatos que não existem. E o pior: eles escrevem essas inverdades de forma muito confiante e bem redigida (Lewis *et al.*, 2020). Na contabilidade, isso é um risco enorme, pois inventar uma regra tributária poderia levar uma empresa a problemas sérios.

1.3 RETRIEVAL-AUGMENTED GENERATION (RAG)

Para resolver o problema das alucinações, pesquisadores criaram a tecnologia chamada Retrieval-Augmented Generation, ou simplesmente RAG (Lewis *et al.*, 2020). Essa técnica funciona como um processo de pesquisa antes da resposta.

Em vez de deixar a IA responder apenas usando a sua memória de longo prazo (onde pode haver erros), o RAG conecta a inteligência artificial a um banco de dados confiável, como arquivos PDF contendo as normas CPC e livros oficiais. Esse processo transforma todo o texto dos livros em números (um processo matemático chamado *embedding*).

Quando um aluno faz uma pergunta, o sistema busca matematicamente qual parágrafo da lei responde àquela dúvida. O sistema recorta esse texto oficial, entrega para a IA e dá a seguinte instrução: “Responda à pergunta do usuário baseando-se apenas neste texto da lei”. Isso reduz significativamente os erros e ajuda o modelo a se manter fiel aos fatos, tornando a IA uma ferramenta muito mais segura para auditoria e estudos (Topsakal; Akinci, 2023).

1.4 TESTES DE LLMS EM EXAMES PROFISSIONAIS

A verificação do conhecimento da IA em exames oficiais tem avançado rápido na literatura acadêmica. O estudo de Eulerich, Masli e Wood (2023) é a principal referência na área de contabilidade. Os autores testaram o ChatGPT em exames como o CPA (para contadores públicos) e o CMA (para contadores gerenciais) nos Estados Unidos.

A pesquisa demonstrou uma evolução impressionante. No começo, as versões mais antigas do ChatGPT reprovavam, acertando cerca de 53% das provas. Porém, com o uso de versões modernas e de técnicas de instruções guiadas (pedindo para a IA explicar passo a passo o que estava fazendo), as taxas de aprovação subiram para mais de 85%, demonstrando uma capacidade superior à da média dos candidatos humanos (Eulerich; Masli; Wood, 2023).

Além de fazer provas, os pesquisadores também começaram a usar a inteligência artificial para corrigir e avaliar documentos. O estudo de Zheng *et al.* (2023) comprovou que os modelos de ponta conseguem atuar como juízes, avaliando o grau de dificuldade de um texto com um nível de precisão semelhante ao de especialistas humanos. Essa técnica, conhecida como LLM-as-a-judge (LLM como juiz), permite que usemos a IA para classificar o nível de dificuldade do Exame de Suficiência.

2 PROCEDIMENTOS METODOLÓGICOS

2.1 DELINEAMENTO DA PESQUISA E ESCOLHA DAS QUESTÕES

Esta pesquisa é de natureza exploratória, descritiva e documental, utilizando uma abordagem quantitativa para analisar a precisão (acurácia) das respostas e a concordância nas classificações feitas pelos modelos (Bardin, 2011; Krippendorff, 2019).

O conjunto de dados da pesquisa engloba as provas do Exame de Suficiência organizadas ao longo de 9 edições consecutivas, do primeiro semestre de 2022 ao primeiro semestre de 2026. Durante esse período, as provas foram elaboradas por duas bancas diferentes: a Consulplan (2022 e 2023) e a Fundação Getúlio Vargas (de 2024 em diante).

Para evitar que a mesma questão fosse analisada mais de uma vez (já que o CFC aplica provas de cores diferentes com as mesmas perguntas misturadas), o estudo utilizou apenas a “Prova Tipo 1” (Branca) de cada edição, resultando em 450 questões objetivas. Em seguida, foram retiradas 16 questões que haviam sido anuladas oficialmente pelas bancas organizadoras. Assim, a amostra final ficou em 434 questões válidas, utilizadas em todas as análises computacionais.

2.2 EXTRAÇÃO DE DADOS E ESTRUTURA DO BANCO DE INFORMAÇÕES

Para que as inteligências artificiais pudessem ler as questões em grande volume, os arquivos originais em PDF fornecidos pelo CFC precisaram ser convertidos para texto. Isso foi feito por meio de roteiros de programação (scripts) que utilizaram a capacidade de leitura visual do modelo gemini-2.5-pro. Essa tecnologia identificou corretamente os enunciados e, mais importante, transformou tabelas complexas (como Balanços Patrimoniais) em texto formatado, sem perder a organização visual original.

Todas as questões, alternativas corretas e informações sobre o ano e a banca organizadora foram armazenadas em um banco de dados seguro chamado CFC.db. Isso garantiu que os testes subsequentes rodassem de forma organizada e padronizada.

2.3 MODELOS UTILIZADOS E CONFIGURAÇÕES

A pesquisa utilizou duas inteligências artificiais desenvolvidas por empresas concorrentes para comparar os desempenhos. Para simular a realidade de instituições de ensino e estudantes que buscam soluções eficientes e de custo acessível, foram escolhidas versões focadas em rapidez e eficiência, e não as versões mais caras e lentas do mercado:

- **Gemini (Google):** Modelo gemini-3.1-flash-lite.
- **GPT (OpenAI):** Modelo gpt-5.4-nano.

Para garantir o rigor científico, a configuração conhecida como “Temperatura” foi ajustada estritamente para zero em todos os testes. Segundo Wang *et al.* (2023), zerar a temperatura tira a aleatoriedade do modelo de linguagem. Isso significa que, se a máquina for submetida à mesma questão dez vezes, ela dará exatamente a mesma resposta dez vezes, permitindo que a pesquisa seja repetida e confirmada por outros cientistas no futuro.

2.4 APLICAÇÃO DA TÉCNICA RAG E TAREFAS DO EXPERIMENTO

Nas tarefas que precisaram consultar a literatura, o banco de pesquisa (RAG) foi alimentado com documentos oficiais e obras clássicas. Quando uma questão envolvia consulta, o sistema lia a pergunta do CFC, procurava na base de dados os cinco trechos mais relevantes sobre o assunto e enviava tudo para o modelo de IA responder com base na leitura (Lewis *et al.*, 2020).

No total, as duas máquinas executaram de forma independente seis tarefas diferentes, nomeadas de T1 a T6:

- **T1 (Resolução sem consulta):** o modelo precisa resolver a questão apenas com o que já sabe, sem pesquisar em nenhum documento externo (zero-shot). Avalia-se a taxa de acerto bruto.
- **T2 (Resolução com RAG):** o modelo resolve a questão recebendo trechos dos livros e leis de apoio.
- **T3 (Avaliação de dificuldade da questão):** atuando como professor, o modelo avalia de 1 a 5 o quão difícil é o conceito cobrado na questão.

- **T4 (Avaliação de cobertura de material):** o modelo avalia de 1 a 5 se os textos teóricos fornecidos pelo RAG foram realmente úteis para achar a resposta da questão.
- **T5 (Aderência às diretrizes nacionais):** o modelo verifica de 1 a 5 o quanto a questão está alinhada com as habilidades e competências exigidas pelo Ministério da Educação nas DCNs.
- **T6 (Classificação por área do conhecimento):** o modelo classifica obrigatoriamente a questão em uma das 9 áreas contábeis clássicas (Geral, Custos, Pública, Gerencial, etc.).

Para analisar se as IAs da Google e da OpenAI concordam entre si nas tarefas de classificação pedagógica (T3, T4, T5 e T6), o estudo utilizou a métrica estatística chamada Kappa de Cohen (Cohen, 1960). Essa métrica desconta os acertos feitos “por sorte”, demonstrando se há uma concordância genuína e técnica entre as avaliações dos modelos.

3 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

3.1 ACURÁCIA NA RESOLUÇÃO (T1) E A COMPARAÇÃO COM OS CANDIDATOS HUMANOS

A análise de resolução das 434 questões contábeis demonstra claramente a superioridade das ferramentas atuais de inteligência artificial diante da prova oficial, conforme demonstrado na TAB. 1. No cenário chamado de zero-shot (T1), em que a máquina resolve a prova apenas usando o conhecimento gravado em seu sistema interno (sem pesquisar livros ou arquivos PDFs na hora), o modelo Gemini obteve uma taxa impressionante de 89,40% de acertos. Na prática, ele respondeu de forma exata a 388 questões.

O modelo desenvolvido pela OpenAI (GPT-5.4-nano) obteve uma precisão de 74,65% (acertando 324 questões). Apesar de ficar numericamente abaixo do Gemini, é preciso lembrar que o Exame de Suficiência exige apenas 50% de acertos para aprovação legal. Isso comprova que até mesmo as versões mais simples e baratas da inteligência artificial (aquelas focadas em economia de dados e velocidade, como as usadas nesta pesquisa) passariam facilmente na prova do conselho.

Para ilustrar o tamanho desse feito tecnológico, é necessário comparar os acertos das máquinas com o desempenho real dos alunos. A TAB. 1 cruza a média histórica oficial de aprovação no Brasil com a pontuação alcançada pelos robôs na mesma prova.

TABELA 1 – Taxas oficiais de Aprovação (Candidatos vs. IA) por Edição (2022-2026)

Edição	Banca Examinadora	Taxa Aprovação Humana (CFC%)	Acurácia Gemini T1 (%)	Acurácia GPT T1 (%)	Múltiplo de Sucesso (Gemini/ Humano)
2022.1	CONSULPLAN	18,0%	95,8%	77,1%	5,3x
2022.2	CONSULPLAN	17,0%	91,5%	63,8%	5,4x
2023.1	CONSULPLAN	13,0%	81,2%	68,8%	6,2x
2023.2	CONSULPLAN	13,0%	82,6%	71,7%	6,4x
2024.1	FGV	48,2%	98,0%	84,0%	2,0x
2024.2	FGV	12,1%	83,7%	69,4%	6,9x
2025.1	FGV	Não divulgado	93,8%	83,3%	-
2025.2	FGV	Não divulgado	91,7%	77,1%	-
2026.1	FGV	Não divulgado	86,0%	76,0%	-

FONTE: Adaptado de CFC (2026)

Os dados da TAB. 1, revelam que o índice de aprovação humana tem oscilado negativamente em grande parte das provas. Na prova 2024.2, do Conselho Federal de contabilidade (CFC), organizada pela Fundação Getúlio Vargas (FGV), a prova derrubou as expectativas de quase 88% dos estudantes nacionais, permitindo uma aprovação real de apenas 12,1%. Exatamente na mesma edição e no mesmo questionário, a inteligência artificial do Google gabaritou 83,7% da prova. Isso significa um desempenho quase 7 vezes superior ao humano na edição mais rígida da banca organizadora.

Percebe-se também que os modelos de linguagem acompanham o nível de dificuldade imposto. Quando a banca facilitou a prova na edição 2024.1 (levando o índice humano a atingir atípicos 48,2%), o Gemini também se beneficiou da clareza das perguntas, alcançando a marca quase perfeita de 98,0% de exatidão. Os resultados no Brasil confirmam conclusões semelhantes apontadas por Eulerich, Masli e Wood (2023) nas pesquisas feitas nos Estados Unidos. Eles demonstraram que exames de perguntas e respostas objetivas, por mais complexos que pareçam, encontram-se hoje inteiramente dominados pelas capacidades lógicas das inteligências gerativas contemporâneas.

3.2 A INJEÇÃO DE CONTEXTO POR RAG (T2) E O PROBLEMA DA ALUCINAÇÃO INDUZIDA

Diante das afirmações teóricas trazidas por pesquisadores de que o acesso livre a uma biblioteca diminui os erros das inteligências (Lewis *et al.*, 2020), o passo metodológico

seguinte submeteu os dois modelos novamente às 434 perguntas (Tarefa 2), mas agora oferecendo acesso eletrônico instantâneo aos textos oficiais e manuais do MEC.

De maneira surpreendente, os dados mostraram um resultado muito tímido para um modelo e uma piora real no outro. O Gemini obteve um aumento de apenas 0,23 pontos, subindo de 89,40% para 89,63%. O uso das normas ajudou a ferramenta a corrigir 17 questões em que havia escorregado na etapa anterior. No entanto, o fluxo excessivo de textos de leis confundiu a interpretação da própria máquina, levando-a a errar de forma induzida 16 questões que ela já havia respondido corretamente antes da ajuda dos textos.

O fenômeno foi muito mais agressivo na inteligência da OpenAI (GPT). Com o uso dos materiais, em vez de melhorar, sua taxa global despencou de 74,65% para 73,04%. Esse fato inesperado levanta um novo conceito dentro do estudo da automação contábil: a “alucinação induzida por ruído” ou over-grounding bias.

Como os assuntos avaliados pelo CFC versam sobre normas internacionais e regras de débito e crédito imutáveis (os Pronunciamentos Técnicos já estão globalmente pacificados), os LLMs de última geração já estudaram todo esse material contábil milhares de vezes durante as suas criações em laboratório (Topsakal; Akinci, 2023). Fornece, de forma insistente, pedaços avulsos de manuais e pronunciamentos na hora da questão faz com que o robô abandone sua intuição matemática lógica em prol da imitação do texto oferecido. Assim, para matérias estritamente lógicas e fechadas da contabilidade, o RAG não se prova benéfico nos níveis técnicos iniciais, ao contrário de matérias abertas em interpretação ou puramente tributárias mutáveis.

3.3 DIFERENÇAS NA AVALIAÇÃO PEDAGÓGICA DE DIFICULDADE (T3 E T4)

Quando testados como se fossem professores responsáveis por julgar a dificuldade das perguntas aplicadas pelo conselho federal, os modelos revelaram pensamentos muito destoantes. Esse experimento buscou validar a ideia descrita no exterior sob a expressão “LLM como juiz” (Zheng *et al.*, 2023).

Para o modelo da Google, as provas contábeis brasileiras transitam amplamente entre os conceitos de “Fácil” e “Médio”. Na pontuação estrita que ia de 1 a 5, o sistema computou a média global como 2,46. Em visão totalmente oposta, o motor da OpenAI avaliou o exame de forma muito estagnada e monótona. Para o GPT, incríveis 93% das questões da história recente caem rigidamente na caixa de nível 3 (Médio).

Ao cruzar os laudos opinativos de dificuldade entregues por cada empresa utilizando o coeficiente Kappa de Cohen (Cohen, 1960), a fórmula apurou um resultado estatístico de míseros 0,0573 (Concordância Pobre). A conclusão lógica para o ambiente

educacional indica que, apesar de resolverem contas e impostos da mesma forma precisa, as máquinas discordam profundamente na hora de categorizar se algo foi ou não um desafio cognitivo humano. Contudo, em uma constatação curiosa de consistência individual, ambas as inteligências consideraram internamente que as (poucas) questões que elas acabaram errando na T1 eram, de fato, estatisticamente mais complexas e difíceis do que aquelas que elas acertaram logo de início.

3.4 DIVERGÊNCIAS NA ADERÊNCIA ÀS DIRETRIZES NACIONAIS CURRICULARES (T5)

Um debate antigo em meio às coordenações acadêmicas universitárias trata do descolamento entre as exigências reais impostas pelas provas anuais do CFC e os eixos temáticos humanizados requeridos pelo próprio Ministério da Educação através das suas Diretrizes Curriculares (Brasil, 2024).

Colocados sob a função de auditar a adequação da prova aos anseios do MEC, o descompasso interpretativo se fez novamente colossal (Kappa negativo de -0,01). A inteligência Gemini assumiu uma visão positiva que chega a flertar com a ilusão institucional, sustentando relatórios indicando que mais de 93% das problemáticas criadas para o exame exalariam aderência formidável e obediência total ao currículo estipulado. Já a lógica programática da OpenAI, assumindo feições interpretativas notadamente desconfiadas, analisou friamente e deduziu que apenas 56,7% do volume provado atende fielmente aos comandos estruturantes exigidos na DCN contábil contemporânea (Padoveze, 2021).

3.5 CLASSIFICAÇÃO TEMÁTICA DAS ÁREAS DE CONHECIMENTO (T6)

Deixando as classificações opinativas e adentrando nos blocos de categorização dura, o teste final pediu que os modelos abrigassem e enquadrassem as 434 questões em uma das 9 rotas disciplinares estipuladas secularmente pelos editais (Geral, Custos, Pública, Perícia, Ética, etc.). Diferentemente da desordem subjetiva anterior, nesta tarefa as máquinas apresentaram um índice esmagador de Concordância Substancial: o Kappa subiu vigorosamente para 0,77.

A harmonia permitiu confirmar onde o conselho de classe efetivamente deposita a âncora da avaliação no aluno nacional:

1. Contabilidade Geral e Societária (48,6% sob a ótica Gemini e 41,5% para o GPT);
2. Ética Profissional e Legislação Normatizada (15,4% Gemini; 18,9% GPT);
3. Contabilidade de Custos (8,5% Gemini; 11,5% GPT).

A junção unificada e consolidada desses três ramos suga implacavelmente entre 70% e 77% da formulação inteira da vida provatória do estudante (CFC, 2024). O impacto brutal dessas evidências estatísticas soa o alarme sobre o abismo educacional brasileiro focado na “Contabilidade Gerencial” (foco da DCN moderna no julgamento da controladoria, previsão de fluxos, formação de lucro e margem variável). Esse viés gerencial agoniza em presenças minúsculas que variam entre apenas 2,5% e 4,1% do conteúdo cobrado na prova, confirmando a tese acadêmica de teóricos como Martins (2023) e Padoveze (2021) de que os avaliadores exigem habilidades rotineiras procedimentais e ignoram em exames genéricos a postura de gestor de lucros exigida pela Resolução CNE/CES nº 1/2024.

CONSIDERAÇÕES FINAIS

O objetivo traçado por esta pesquisa foi atestar, em bases matemáticas rígidas, o patamar operacional da inteligência artificial generativa em resolver exames profissionais altamente seletivos no Brasil. Sob testes de simulação determinística sem aleatoriedades (garantidos pelo bloqueio da temperatura algorítmica), buscou-se também forçar o sistema a assumir a identidade didática e validar, ou não, as diretrizes propostas pelo Ministério da Educação.

Em resposta direta à pergunta norteadora de investigação: sim, os modelos de linguagem arquitetados recentemente estão amplamente capacitados a alcançar as marcas legais para aprovação com margens elásticas e confortáveis. O desempenho do construto gemini-3.1-flash-lite, alcançando uma exatidão puramente textual de 89,4% na grande amostra contábil validada e transpondo com louvor os abismos das transições entre as bancas avaliadoras do exame do conselho Federal de Contabilidade, a Fundação Getulio Vargas, e o Instituto Consulplan, endossa amplamente no hemisfério sul os estudos pioneiros de Eulerich, Masli e Wood (2023), humilhando estatisticamente o modesto retrospecto contemporâneo fixado pelos universitários brasileiros (que resvala em sucessivas aprovações na casa dos pífios 12% aos 18%).

Uma descoberta contundente e técnica foi derivada do experimento de incremento textual pela engenharia RAG. A provisão em tempo real do livro eletrônico normativo inserida no questionário piorou efetivamente a tomada de decisão do modelo da OpenAI de forma sensível. Conclui-se que o ferramental de ponta comercializado já carrega o corpo da lei consolidada da profissão imantado em suas conexões centrais de aprendizagem e, por isso, injetar excertos bibliotecários limitantes induz o cérebro sintético à desconcentração lógica nas exatas perguntas mecanicistas, provocando uma ineficiência nominada no estudo de alucinação por indução de ruído.

A repercussão final das informações apuradas estende pontes para duas províncias urgentes e conflitantes (Becker, 1964). Nas academias de formação, os dados corroboram com entusiasmo a introdução veloz dessas arquiteturas robóticas trabalhando em uníssono como preceptores individualizados gratuitos na rotina do acadêmico rumo ao CFC. Constatou-se altíssima concordância ($Kappa = 0,77$) da máquina no ato de classificar, engavetar e separar por área específica os certames dispostos nos editais, validando a inteligência não só na lide das soluções prontas, mas na separação temática robusta.

No âmbito da regulação oficial e preservação mercadológica amparada no conceito de sinalização (Spence, 1973), soa um forte aviso sonoro à autarquia federal responsável pelas matrizes. Constatou-se inequivocamente que a prova não acompanha os ensejos de maturidade interpretativa exarados pelas atuais DCNs (Brasil, 2024). Os resultados atestam uma saturação conteudista e avaliativa enraizada quase em 80% sobre os velhos dogmas fixos e procedimentos herméticos da Contabilidade Geral. Essa é uma competência mecânica pura em que a automação computacional hoje não apenas habita, mas domina integralmente de olhos fechados. Exige-se das bancas a confecção impeditiva focada em resolução aberta, dilemas gerenciais flexíveis e periciais éticas difusas para se medir de fato um analista corporativo que faça jus ao século tecnológico.

Foi observado, em relação ao resultado de exatidão das IAs utilizadas, em específico o Gemini, que foi de 89,4% a dificuldade de interpretação nas questões, demonstrada como incerteza de dar uma resposta correta, justificando o percentual de acerto. Isso pode ser interpretado como erros na formulação das questões pelas bancas avaliadoras que compromete a compreensão das questões pelos alunos que realizaram a prova.

Portanto, o trabalho esbarrou em provisões limitantes pela exclusão forçosa de material dependente imperiosamente de diagramação imagética que não pôde ser transportada para linguagem legível em textos no formato de tabela. Recomenda-se para o futuro a emulação avaliativa das edições contábeis mediante os gigantescos modelos visuais e multimodais fechados (as versões Advanced das desenvolvedoras), checando em bases plurais se a baixa sintonia opinativa apurada nas matrizes educacionais (dificuldade e aderência à DCN) continua a gerar ruídos metodológicos profundos nas grandes frentes de pesquisa tecnológica.

REFERÊNCIAS

- BARDIN, L. **Análise de conteúdo**. São Paulo: Edições 70, 2011.
- BECKER, G. S. **Human capital**: A theoretical and empirical analysis, with special reference to education. New York: Columbia University Press, 1964.
- BRASIL. Ministério da Educação. **Resolução CNE/CES nº 10, de 16 de dezembro de 2004**. Institui as Diretrizes Curriculares Nacionais (DCN) do curso de graduação em Ciências Contábeis. Diário Oficial da União, Brasília, DF, 17 dez. 2004.
- BRASIL. Presidência da República. **Lei nº 12.249, de 11 de junho de 2010**. Altera a legislação tributária e institui o Exame de Suficiência. Diário Oficial da União, Brasília, DF, 14 jun. 2010.
- BRASIL. Ministério da Educação. **Resolução CNE/CES nº 1, de 31 de janeiro de 2024**. Institui as novas Diretrizes Curriculares Nacionais do curso de graduação em Ciências Contábeis (DCN). Diário Oficial da União, Brasília, DF, 02 fev. 2024.
- COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, [s.l.], v. 20, n. 1, p. 37-46, 1960.
- CONSELHO FEDERAL DE CONTABILIDADE (CFC). **Exame de Suficiência**: Relatórios Estatísticos e Resultados. Brasília, DF: CFC, 2024. Disponível em: <https://cfc.org.br/>. Acesso em: 05 jul. 2026.
- CONSELHO FEDERAL DE CONTABILIDADE (CFC). **Exame de Suficiência**: edições anteriores. Brasília: CFC, 2026. Disponível em: <https://cfc.org.br/exame-de-suficiencia-antiores/2o-exame-de-suficiencia-de-2025/>. Acesso em: 23 set. 2026.
- EULERICH, M.; MASLI, A.; WOOD, D. A. Can Artificial Intelligence Pass Accounting Certification Exams? ChatGPT: CPA, CMA, CIA, and EA? **SSRN Electronic Journal**, [s.l.], 2023. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4440058. Acesso em: 05 jul. 2026.
- FUNDAÇÃO INSTITUTO DE PESQUISAS CONTÁBEIS, ATUARIAIS E FINANCEIRAS (FIPECAFI). **Manual de contabilidade societária**: aplicável a todas as sociedades de acordo com as normas internacionais e do CPC. 4. ed. São Paulo: Atlas, 2022.
- KRIPPENDORFF, Klaus. **Content analysis**: An introduction to its methodology. 4th ed. Thousand Oaks: SAGE Publications, 2019.
- LEWIS, P. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. **Advances in Neural Information Processing Systems**, [s.l.], v. 33, p. 9459-9474, 2020.
- MARTINS, E. **Contabilidade de custos**. 12. ed. São Paulo: Atlas, 2023.
- PADOVEZE, C. L. **Contabilidade gerencial**: um enfoque em sistema de informação contábil. 9. ed. São Paulo: Atlas, 2021.
- PELEIAS, I. R. *et al.* **Evolução do ensino da contabilidade no Brasil**: uma análise histórica. Revista Contabilidade & Finanças, São Paulo, v. 18, n. 43, p. 19-32, 2007.
- SCHULTZ, Theodore W. Investment in Human Capital. **The American Economic Review**, [s. l.], v. 51, n. 1, p. 1-17, 1961.
- SPENCE, M. Job market signaling. **The Quarterly Journal of Economics**, [s. l.], v. 87, n. 3, p. 355-374, 1973.

TOPSAKAL, Oguzhan; AKINCI, Tahir Cetin. Creating Large Language Model Applications Utilizing LangChain: a primer on developing LLM apps fast. In: INTERNATIONAL CONFERENCE ON APPLIED ENGINEERING AND NATURAL SCIENCES, 1., 2023, [S. l.]. **Anais eletrônicos...** [S. l.]: ICAENS, 2023. p. 1050-1056. DOI: <https://doi.org/10.59287/icaens.1127>. Disponível em: <https://as-proceeding.com/index.php/icaens/article/view/1127>. Acesso em: 05 jul.2026.

WANG, Y. *et al.* Reproducibility in Large Language Models: a study of temperature settings. [S. l.]: **arXiv**, 2023. arXiv:2310.01234.

ZHENG, L. *et al.* Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 37., 2023, New Orleans. **Proceedings... Red Hook**, NY: Curran Associates / ACM, 2023. art. 2020, p. 46595-46623. Disponível em: <https://dl.acm.org/doi/10.5555/3666122.3668142>. Acesso em: 5 out. 2026.