

APLICAÇÃO DE INTELIGÊNCIA ARTIFICIAL NO PROCESSO AVALIATIVO DE ALUNOS DE UMA EMPRESA DO SETOR EDUCACIONAL

1º ANTONIO CARLOS PRESTES PERINE - pg405166@uem.br
UNIVERSIDADE ESTADUAL DE MARINGÁ - UEM

2º LARISSA CAROLINE DE BRITO - pg407154@uem.br
UNIVERSIDADE ESTADUAL DE MARINGÁ - UEM

3º HEITOR DELIVIO SANCHES SALINAS - pg407153@uem.br
UNIVERSIDADE ESTADUAL DE MARINGÁ - UEM

4º FRANCIELLE CRISTINA FENERICH- fcfenerich@uem.br
UNIVERSIDADE ESTADUAL DE MARINGÁ - UEM

5º MARCO ANTONIO FERREIRA - marcoferreira@utfpr.edu.br
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ - UTFPR

ÁREA: *6. GESTÃO ORGANIZACIONAL*
SUBÁREA: *6.8. GESTÃO DO CONHECIMENTO*

RESUMO*: A ADOÇÃO DE TECNOLOGIAS DE INTELIGÊNCIA ARTIFICIAL (IA) NO AMBIENTE EDUCACIONAL REVOLUCIONOU NÃO APENAS A ENTREGA DE CONTEÚDO, MAS CRUCIALMENTE OS MÉTODOS DE AVALIAÇÃO DOS ALUNOS. ESTE ESTUDO EXPLORA A EFICÁCIA DA IA GENERATIVA, ESPECIFICAMENTE DO CHATGPT 4.0, NA AVALIAÇÃO DE RESPOSTAS SUBJETIVAS NO ENSINO SUPERIOR. UTILIZANDO UMA ESCALA LIKERT PARA MEDIR A PRECISÃO E A COMPLETUDE, A PESQUISA COMPARA AS AVALIAÇÕES GERADAS PELA IA COM AQUELAS FEITAS POR EDUCADORES HUMANOS. OS RESULTADOS INDICAM QUE O CHATGPT 4.0, SEM TREINAMENTO PRÉVIO, MUITAS VEZES FORNECE AVALIAÇÕES COMPARÁVEIS ÀS DOS EDUCADORES, ENQUANTO O TREINAMENTO ADICIONAL DA IA PODE LEVAR A UMA REDUÇÃO NA PRECISÃO. ESSES ACHADOS DESTACAM O POTENCIAL E AS LIMITAÇÕES DA IA GENERATIVA NA AVALIAÇÃO EDUCACIONAL, ENFATIZANDO A NECESSIDADE DE CALIBRAÇÃO CUIDADOSA E COMPREENSÃO DAS CAPACIDADES DA IA PARA APRIMORAR A EXPERIÊNCIA DE APRENDIZAGEM.*

Keywords: *INTELIGÊNCIA ARTIFICIAL; ENSINO SUPERIOR; IA GENERATIVA; CHATGPT; ESCALA LIKERT*

APPLICATION OF ARTIFICIAL INTELLIGENCE IN THE STUDENT ASSESSMENT PROCESS IN AN EDUCATIONAL SECTOR COMPANY

ABSTRACT: *THE ADOPTION OF ARTIFICIAL INTELLIGENCE (AI) TECHNOLOGIES IN EDUCATIONAL SETTINGS HAS REVOLUTIONIZED NOT ONLY CONTENT DELIVERY BUT ALSO CRUCIALLY THE METHODS OF STUDENT ASSESSMENT. THIS STUDY EXPLORES THE EFFECTIVENESS OF GENERATIVE AI, SPECIFICALLY CHATGPT 4.0, IN EVALUATING SUBJECTIVE RESPONSES WITHIN HIGHER EDUCATION. UTILIZING A LIKERT SCALE TO MEASURE ACCURACY AND COMPLETENESS, THE RESEARCH COMPARES AI-GENERATED ASSESSMENTS AGAINST THOSE MADE BY HUMAN EDUCATORS. RESULTS INDICATE THAT CHATGPT 4.0, WITHOUT PRIOR TRAINING, OFTEN PROVIDES ASSESSMENTS COMPARABLE TO THOSE OF EDUCATORS, WHILE ADDITIONAL TRAINING OF THE AI CAN LEAD TO REDUCED ACCURACY. THESE FINDINGS HIGHLIGHT THE POTENTIAL AND LIMITATIONS OF GENERATIVE AI IN EDUCATIONAL EVALUATION, EMPHASIZING THE NEED FOR CAREFUL CALIBRATION AND UNDERSTANDING OF AI CAPABILITIES TO ENHANCE THE LEARNING EXPERIENCE.*

Keywords: *ARTIFICIAL INTELLIGENCE; HIGHER EDUCATION; GENERATIVE AI; CHATGPT; LIKERT SCALE*

.

1. INTRODUÇÃO

A adoção de tecnologias de Inteligência Artificial (IA) no ambiente educacional tem se mostrado uma revolução não apenas na entrega de conteúdo, mas crucialmente nos métodos de avaliação de alunos. Em universidades privadas, onde a busca por inovação e excelência é constante, a IA surge como uma ferramenta potencialmente transformadora. Segundo o "Consenso de Beijing sobre IA e Educação" (2019), enfatiza-se a importância de integrar a IA para fortalecer políticas educacionais e promover a aprendizagem ao longo da vida, alinhado ao Objetivo de Desenvolvimento Sustentável número 4 da UNESCO. Este enfoque está em sintonia com a necessidade de uma avaliação mais profunda e personalizada, que não apenas mede, mas também orienta o desenvolvimento educacional dos alunos (Kizilcec, 2023).

O processo avaliativo, essencial para entender tanto o progresso quanto a profundidade da compreensão dos alunos, evolui com o uso de IA. Através de ferramentas como Processamento de Linguagem Natural (PLN), a avaliação transcende a verificação baseada em palavras-chave para uma análise contextual significativa das submissões dos alunos. Esta capacidade não só promove um feedback mais imediato e relevante, mas também refinar continuamente os métodos de ensino e aprendizagem nas universidades (Baker et al., 2020).

Contudo, a implementação da IA no processo avaliativo carrega desafios significativos. Questões de equidade, privacidade dos dados e precisão dos resultados demandam uma consideração ética e prática rigorosa. A resistência de educadores e alunos, muitas vezes devido a uma falta de compreensão das capacidades e limitações da IA, pode impedir sua adoção efetiva. É crucial, portanto, que as instituições não somente adotem essas tecnologias, mas também invistam na educação de seu corpo docente e administrativo sobre o funcionamento e as aplicações éticas da IA (Arnold & Pistilli, 2012; Floridi & Cowls, 2022).

Finalmente, é importante notar que enquanto ferramentas como o ChatGPT oferecem amplas possibilidades para a transformação dos processos educacionais, elas devem ser utilizadas com um entendimento crítico de suas limitações. O potencial de tais ferramentas para redefinir os paradigmas educacionais é imenso, contudo, deve-se proceder com cautela para assegurar que a inovação tecnológica contribua positivamente para o ambiente educacional e não comprometa a integridade acadêmica ou a segurança dos dados dos alunos (Ribera & Díaz Montesdeoca, 2024).

Este artigo visa apurar a acurácia da aplicação de IA generativa no contexto de correção de respostas subjetivas em avaliações de alunos, para isso, utilizou-se métodos estatísticos como,

teste de Friedman para verificar se há diferenças entre os grupos de correções (com ou sem treinamento prévio da IA) e Nemenyi com o intuito identificar quais pares de métodos de correção são significativamente diferentes entre si, a fim de avaliar a capacidade das IA generativas de julgar e validar respostas subjetivas no contexto educacional e sua eficácia prática. Propõe-se também uma metodologia de avaliação de alunos por meio de um questionário online, com posterior transformação dos dados qualitativos em quantitativos para uso da escala likert. O estudo explora o uso de IA generativa na educação e introduz uma abordagem inovadora para a avaliação educacional.

2. REVISÃO BIBLIOGRÁFICA

2.1 O uso da inteligência artificial no meio educacional

Assim como citado anteriormente, segundo a UNESCO (2020), diversos especialistas em IA participaram da Conferência de Beijing e concordaram com o "Consenso de Beijing sobre IA e Educação". Neste documento, destacam a relevância de promover a IA na educação, conforme o Objetivo de Desenvolvimento Sustentável número 4 estabelecido pela UNESCO. Sendo assim, as principais ações a serem promovidas são: planejar políticas educacionais, promover usos para gestão educacional, capacitar professores e aprendizes, promover valores e habilidades para a vida e o trabalho, oferecer oportunidades de aprendizagem ao longo da vida e romper as barreiras no mundo digital, incentivar usos equitativos e inclusivos, incluindo igualdade de gênero; usos éticos e transparentes; pesquisar, avaliar e apoiar a inovação.

Diversas experiências reais interessantes em torno da IA são compiladas no Relatório Horizon (CHEW; CHUA, 2020). Exemplos incluem o Watson Tutor da IBM, testado com sucesso em escolas no Texas e no Canvas LMS, o DSMLP da UC San Diego, que oferece acesso a recursos e projetos estudantis através de aprendizado de máquina, o Boost, um aplicativo móvel integrado ao Canvas LMS para assistência no aprendizado online, e o Edulai, uma ferramenta inteligente para avaliação e desenvolvimento de habilidades sociais. Os especialistas consideram que, apesar das preocupações éticas, as aplicações de IA na educação estão previstas para um crescimento significativo (CHEW; CHUA, 2020, p. 27).

Outro relatório relevante da UNESCO (2020) analisa o uso de IA em três áreas: aprendizado com IA, aprendizado sobre IA e preparando-se para IA, com foco no potencial dessas tecnologias para transformar vidas. Experiências educacionais com IA geralmente estão relacionadas ao tutoramento ou à avaliação.

Segundo Mireia Ribera Oliver Díaz Montesdeoca (2024), a inteligência artificial no meio educacional é usada como assistente, inspiração, corretor linguístico, para transformar textos, iterar e refinar ideias e para tarefas repetitivas. A capacidade de raciocinar linguísticamente do ChatGPT deve ser equilibrada com sua falta de conhecimento do mundo real, o que pode levar à apresentação de fatos falsos ou inventados. É importante tomar suas recomendações com um olhar crítico para evitar possíveis perigos. O ChatGPT é uma ferramenta de domínio genérico, não especializada em uma disciplina específica, com uma base de conhecimento geral. Portanto, em áreas especializadas como medicina ou arquitetura, pode ser útil procurar alternativas ou complementar com ferramentas específicas desse domínio.

Consoante a Montesdeoca (2024) por outro lado, a OpenAI é uma empresa comercial que oferece essa ferramenta em nível experimental. Ao concordarmos em usá-la, estamos compartilhando as informações fornecidas como base de treinamento. É prudente não compartilhar informações privadas que possam afetar a segurança dos sistemas informáticos ao interagir com o ChatGPT. Ao conhecer seus limites e aproveitar seu potencial, o ChatGPT representa uma mudança fundamental nos processos de comunicação e na criação de conhecimento, com um efeito disruptivo na educação. Os efeitos do ChatGPT na educação são muito impressionantes. Assim como a irrupção das enciclopédias online questionou a importância de memorização de muitos dados, o surgimento do ChatGPT levanta a necessidade de repensar onde está o valor da aprendizagem. Talvez com IA generativa, não faz mais sentido criar conteúdo repetidamente, devendo nos concentrar principalmente na criatividade e na capacidade de especificar que tipo de texto, imagem ou vídeo precisamos em cada oportunidade. O desenvolvimento do pensamento crítico também parece ser essencial.

2.2 Avaliação de Alunos com IA e Feedback

Alguns estudiosos exploraram o uso da IA para avaliação de alunos e feedback a respeito de performances, abaixo segue alguns tópicos abordados.

Mirchi, Bissonnette, Yilmaz, Ledwos, Winkler-Schwartz e Del Maestro (2020) utilizaram IA para criar um Assistente Operativo Virtual na área da medicina, este assistente fornece feedback automático aos alunos com base em métricas de desempenho, integrando realidade virtual e IA para melhorar o aprendizado. No campo do e-learning na medicina, Janpla e Piriyasurawong (2020) desenvolveram um software inteligente que produz testes online. Este software usa IA para selecionar questões e avaliar os alunos.

No campo do feedback, alguns autores trouxeram pontos interessantes, Saplacan, Herstad, Pajalic (2018) analisaram o feedback fornecido por sistemas digitais em situações de aprendizagem. Eles identificaram problemas como emoções negativas nos alunos e propuseram que as interfaces digitais também despertem emoções positivas Samarakou, Fylladitakis, Karolidis, Früh, Hatziapostolou, Athinaios, Grigoriadou (2016) focaram no monitoramento contínuo e avaliação de estudantes de engenharia. Seu sistema, chamado StuDiAsE, baseado em IA, proporciona feedback personalizado e avaliação de desempenho com informações quantitativas e qualitativas.

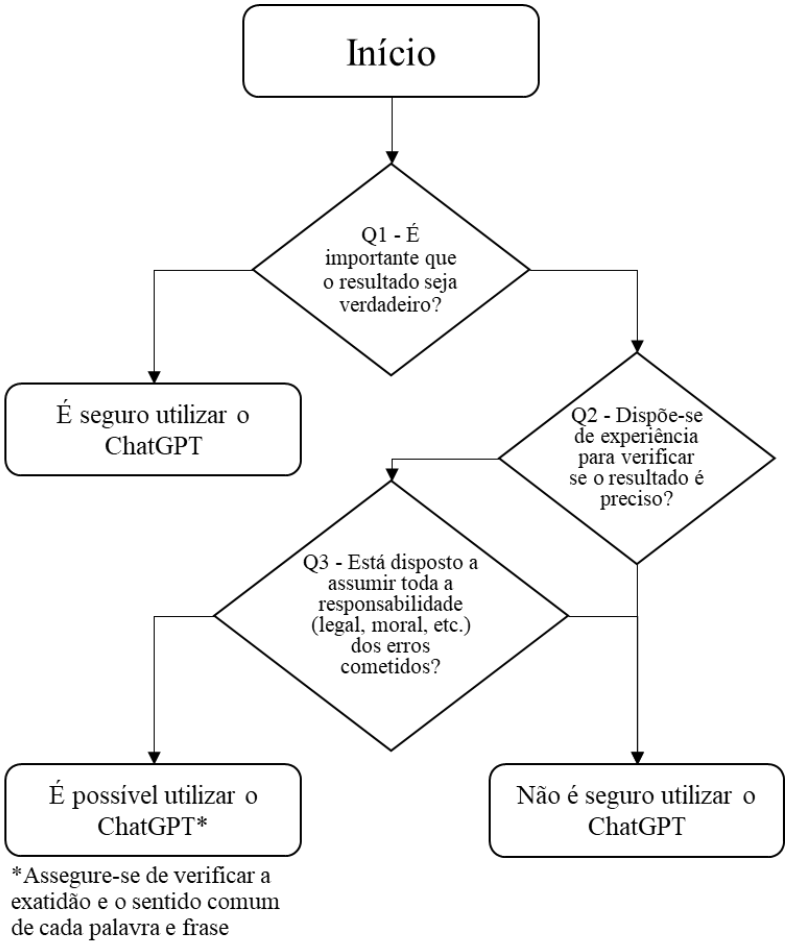
Outro método interessante foi destacado por Rodríguez-Ascaso, Boticario, Finat e Petrie (2017) exploraram sistemas de aprendizagem adaptativa e autoavaliação. Eles desenvolveram um sistema personalizado que permite aos alunos, inclusive aqueles com deficiências, autoavaliar-se e expressar suas preferências para acessar materiais de aprendizagem eletrônicos.

Esses estudos demonstram a aplicação da IA na avaliação de alunos, desde a criação de assistentes virtuais para feedback até a adaptação de sistemas para diferentes necessidades de aprendizado.

3. METODOLOGIA

Existem considerações importantes ao se utilizar IA generativas na gestão da educação. Sabzalieva e Valentini (2023) trazem um fluxograma de decisões, elaborado por Tiulkanov (2023), que indica quando é seguro utilizar IA generativas, mais especificamente o ChatGPT. Dentro do contexto da educação, o educador fica responsável pela segunda decisão (Q2) que é pautada na disposição de experiência para verificar se o resultado é preciso ou não.

FIGURA 1 - Fluxograma de decisões sobre quando é seguro utilizar o ChatGPT.



Fonte: Traduzido e adaptado de Sabzalieva e Valentini (2023) elaborado por Tiulkanov (2023).

Com o propósito de investigar se IA generativas, mais especificamente o ChatGPT, podem assumir o papel do educador dentro do contexto de correção de provas, atividades e exercícios, foi realizado um estudo de caso dentro de uma grande universidade do Brasil seguindo a metodologia descrita na TABELA 1.

TABELA 1 - Metodologia proposta para avaliação de IA generativas no contexto de correção de questões.

Caso	Tipos de questões avaliadas	IA generativas utilizadas para as correções	Treinamento prévio do ChatGPT	Detalhes	Escala Likert para a acurácia de resposta	Escala Likert para o completude da resposta
Base	Discursivas	N/A	N/A	Correção realizada pelo Educador	1-6	1-3
Caso 1	Discursivas	ChatGPT 4	Sem treinamento prévio	Avaliar a capacidade da ferramenta de realizar a correção utilizando o próprio conhecimento	1-6	1-3
Caso 2	Discursivas	ChatGPT 4	Com treinamento prévio	Avaliar a capacidade da ferramenta de adaptação utilizando o treinamento prévio do educador	1-6	1-3

Fonte: Autores.

Duas escalas primárias foram utilizadas para a avaliação das respostas dadas pelos alunos: uma escala de acurácia e uma escala de conteúdo. Escala de acurácia: foi implementadas uma escala Likert de 6 pontos para avaliar a acurácia de cada respostas (1: completamente incorreta, 2: maior parte incorreta, 3: partes iguais correta e incorreta, 4: mais correta do que incorreta, 5: quase inteiramente correta, 6: correta). Avaliação de Completude: uma escala Likert de três pontos foi utilizada para avaliar a completude de cada resposta (1: incompleta, aborda alguns aspectos da pergunta, mas perde partes significativas ou pontos, 2: adequada, aborda todos os aspectos da pergunta e fornece as informações mínimas necessárias, 3: abrangente, aborda todos os aspectos da pergunta e fornece informações adicionais ou contexto além das expectativas). Esta escala detalhada permitiu uma compreensão do desempenho do modelo de IA. Com base no resultado das medidas, buscou-se correlacionar com os casos de base que foram realizados pelos próprios educadores, a fim de entender se é possível utilizar IA generativa como ferramenta auxiliar de correção de questões. Além disso, os dois casos propostos objetivam o compreendimento da possibilidade de alcançar scores mais fiéis aos casos base em diferentes versões do ChatGPT e com a introdução de um treinamento prévio realizado pelo educador para guiar a IA generativa para avaliações mais precisas. O treinamento prévio foi realizado fornecendo informações relevantes para a IA generativa que poderiam ser utilizadas como base de conhecimento, caso dentro do próprio banco de dados não houvesse informações suficientes para a correção da questão.

Neste contexto, foi estruturada uma avaliação, junto ao corpo docente da instituição de ensino em que foram avaliadas 5 questões discursivas e 10 questões objetivas de 54 alunos do curso superior de Análise e Desenvolvimento de Sistemas da disciplina de Go - Projeto de Vida de uma grande universidade privada do Brasil. Para construção do processo avaliativo, foi utilizado a figura 1 para tomadas de decisões e a tabela 1 para entender como seria utilizado a ferramenta IA com maior eficiência.

4. RESULTADOS

Após aplicação da avaliação construída pelo ChatGPT 4.0 e a correção das respostas utilizando a mesma ferramenta de IA, foram conduzidas avaliações estatísticas neste estudo a fim de identificar se existe diferença significativa entre a correção das respostas das questões discursivas entre o educador, a IA generativa (ChatGPT 4.0) sem treinamento prévio e a IA generativa (ChatGPT 4.0) com treinamento prévio. Todos os dados coletados foram sumarizados utilizando métodos estatísticos descritivos. Esses incluíram a média, desvio padrão, mediana e variância. Dada a natureza dos dados, foram utilizados testes estatísticos não-paramétricos para realizar a estatística inferencial. O teste de Friedman foi utilizado para verificar se há diferenças significativas entre múltiplos grupos, avaliando se as classificações médias dos grupos são diferentes. Se o resultado do teste de Friedman for estatisticamente significativo (ou seja, se o valor-p for menor que um nível de significância pré-definido, como 0.05), isso indica que há pelo menos uma diferença significativa entre os métodos de correção. Caso o teste de Friedman ainda indique que há diferenças significativas, para avaliar individualmente cada grupo, realizamos testes post-hoc para identificar quais pares de métodos de correção são significativamente diferentes entre si. Para tanto, escolhemos o teste de Nemenyi para comparações múltiplas. As análises estatísticas foram realizadas usando programação em Python.

TABELA 2 - Valores estatísticos e P-valores para as variáveis de acurácia de completude na correção das 5 questões discursivas e do total que engloba a avaliação de todas as questões em conjunto.

Variáveis	n	Base	Caso 1	Caso 2	valor-P		
					Friedman	Post-hoc - Nemenyi	
					k = 3	Base vs. Caso 1	Base vs. Caso 2

Questão 1: Acurácia	54	3.6±1.2 (4; 1.5) [2.2]	4.3±1.0 (5; 1.1) [2.9]	2.5±1.2 (2; 1.4) [1.2]	1.5E-14	0.03	0.00
Questão 1: Completude	54	1.4±0.6 (1; 0.4) [1.8]	2.0±0.6 (2; 0.4) [2.9]	1.6±0.7 (1; 0.5) [2.1]	8.7E-08	0.00	0.42
Questão 2: Acurácia	54	3.6±0.9 (3; 0.8) [2.1]	4.1±1.0 (4; 1.1) [2.8]	3.3±0.8 (3; 0.7) [1.7]	4.5E-07	0.03	0.31
Questão 2: Completude	54	1.7±0.5 (2; 0.3) [1.8]	2.2±0.6 (2; 0.4) [2.7]	2.2±0.6 (2; 0.3) [2.7]	7.1E-08	0.00	0.00
Questão 3: Acurácia	54	3.7±1.3 (4; 1.6) [2.7]	3.0±0.8 (3; 0.6) [1.9]	2.8±0.8 (3; 0.6) [1.6]	3.6E-07	0.00	0.00
Questão 3: Completude	54	1.5±0.6 (1; 0.4) [2.0]	1.9±0.5 (2; 0.2) [2.7]	1.7±0.5 (2; 0.2) [2.4]	4.4E-04	0.00	0.09
Questão 4: Acurácia	54	3.7±1.0 (4; 1.0) [2.4]	3.8±1.3 (4; 1.6) [2.5]	2.7±1.4 (3; 1.9) [1.5]	1.8E-08	0.91	0.00
Questão 4: Completude	54	1.4±0.5 (1; 0.2) [1.7]	2.1±0.8 (2; 0.6) [2.9]	1.8±0.7 (2; 0.4) [2.2]	1.8E-09	0.00	0.03
Questão 5: Acurácia	54	3.8±1.3 (4; 1.6) [2.2]	3.8±1.4 (4; 2.0) [2.3]	3.2±1.4 (3; 1.9) [1.6]	1.3E-03	0.97	0.06
Questão 5: Completude	54	1.6±0.5 (2; 0.3) [1.7]	2.3±0.8 (3; 0.7) [2.7]	2.0±0.9 (2; 0.8) [2.3]	2.0E-07	0.00	0.03
Total: Acurácia	**270**	**3.7±1.1 (4; 1.3) [2.3]**	**3.8±1.2 (4; 1.5) [2.5]**	**2.9±1.2 (3; 1.4) [1.5]**	**5.3E-29**	**0.81**	**0.00**
Total: Completude	**270**	**1.5±0.6 (1; 0.3) [1.8]**	**2.1±0.7 (2; 0.5) [2.8]**	**1.9±0.7 (2; 0.5) [2.4]**	**4.0E-31**	**0.00**	**0.00**

Cada valor na TABELA 2 pode ser compreendido como média±desvio padrão (mediana; variância) [média de ranking de Friedman]. Esses valores foram fornecidos para cada uma das questões e para cada um dos sistemas de correção: Base com a correção do educador, Caso 1 com a correção do ChatGPT 4.0 sem treinamento prévio e Caso 2 com a correção do ChatGPT 4.0 com treinamento prévio. As análises foram conduzidas usando o teste não paramétrico de Friedman para a comparação de todos os grupos com posterior avaliação post-hoc Nemenyi para a identificação dos grupos com diferente significância.

Em relação ao julgamento da acurácia na resposta das questões dissertativas pelos três diferentes grupos:

- Para a questão 1, o Caso 1 teve um julgamento da acurácia significativamente maior que Base e Caso 2. Já o Caso 2 teve um julgamento acurácia significativamente menor que Base e Caso 1.
- Para a questão 2, o Caso 1 teve um julgamento da acurácia significativamente maior que Base. Além disso, a Base não julgou significativamente diferente de Caso 2.
- Para a questão 3, a Base teve um julgamento da acurácia significativamente maior que Caso 1 e Caso 2.
- Para a questão 4, a Base e o Caso 1 não julgaram significativamente diferente. Já o Caso 2, teve um julgamento da acurácia significativamente menor que Base e Caso 1.

- Para a questão 5, a Base e Caso 1 não julgaram significativamente diferente. Além disso, o Caso 2 teve uma tendência de ter um julgamento da acurácia menor que a Base e o Caso 1, mas essa diferença não foi fortemente significativa.
- Para todo o conjunto das questões, a Base e o Caso 1 não julgaram significativamente diferente. Já o Caso 2 teve um julgamento da acurácia significativamente menor que a Base e o Caso 1.

De maneira geral, o método de correção do professor (Base) e o ChatGPT 4.0 sem treinamento (Caso 1) não são significativamente diferentes na maioria das questões. O ChatGPT 4.0 com treinamento (Caso 2) frequentemente tem um julgamento da acurácia significativamente menor do que os outros métodos.

Em relação ao julgamento da completude na resposta das questões dissertativas pelos três diferentes grupos:

- Para a questão 1, Caso 1 teve um julgamento da completude significativamente maior que a Base. Além disso, o julgamento da Base não foi significativamente diferente do Caso 2.
- Para a questão 2, o Caso 1 e o Caso 2 tiveram um julgamento da completude significativamente maior que a Base.
- Para a questão 3, o Caso 1 teve um julgamento da completude significativamente maior que a Base. Além disso, o julgamento da Base não foi significativamente diferente do Caso 2.
- Para a questão 4, o Caso 1 teve um julgamento da completude significativamente maior que Base. Já a Base teve um julgamento da completude significativamente menor que o Caso 2.
- Para a questão 5, o Caso 1 teve um julgamento da completude significativamente maior que a Base. Já a Base teve um julgamento da completude significativamente menor que o Caso 2.
- Para todo o conjunto das questões, o Caso 1 e o Caso 2 tiveram um julgamento da completude significativamente maior que a Base. Já o Caso 1 teve um julgamento da completude significativamente maior que o Caso 2.

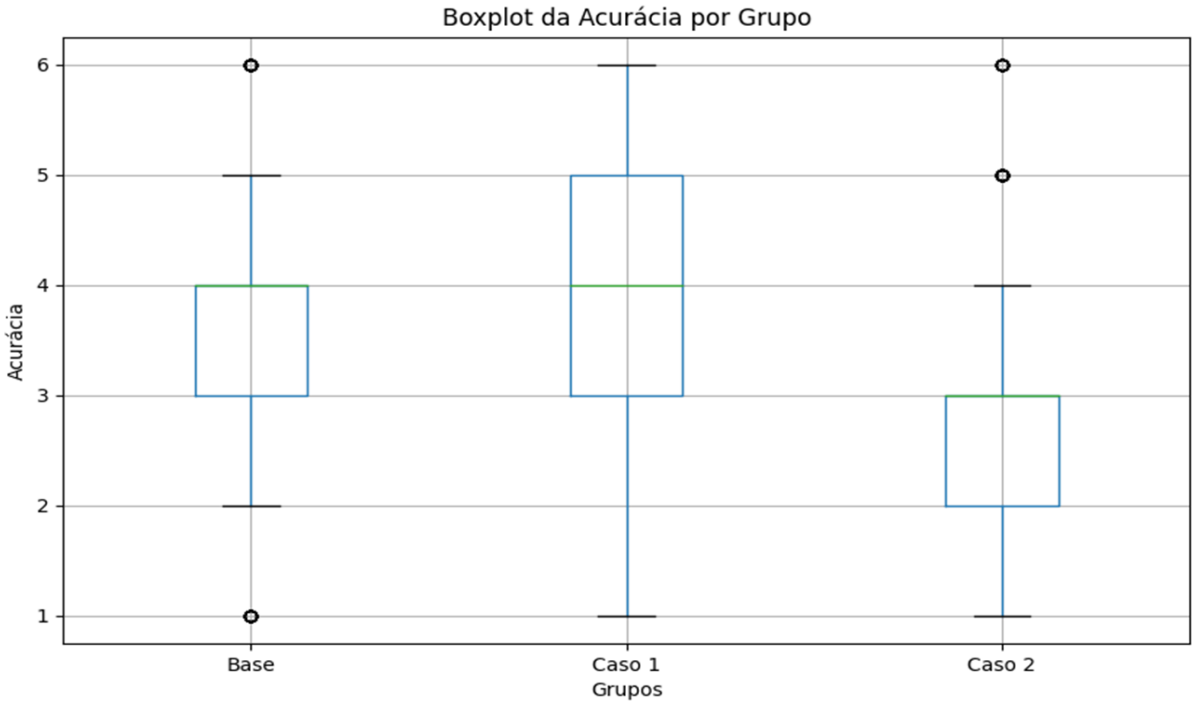
O ChatGPT 4.0 sem treinamento (Caso 1) geralmente tem um julgamento da completude maior, seguido pelo ChatGPT 4.0 com treinamento (Caso 2), e o método de correção do professor (Base) tem o menor julgamento da completude na maioria das questões.

Os resultados da avaliação mostram que a correção feita pelo ChatGPT 4.0 sem treinamento prévio (Caso 1) é frequentemente comparável à correção realizada pelo professor (Base) em termos de acurácia, sem diferenças significativas na maioria das questões. Em contraste, a

versão do ChatGPT 4.0 com treinamento prévio (Caso 2) apresentou um julgamento da acurácia significativamente menor em várias questões, indicando que o treinamento adicional não melhorou a precisão das correções.

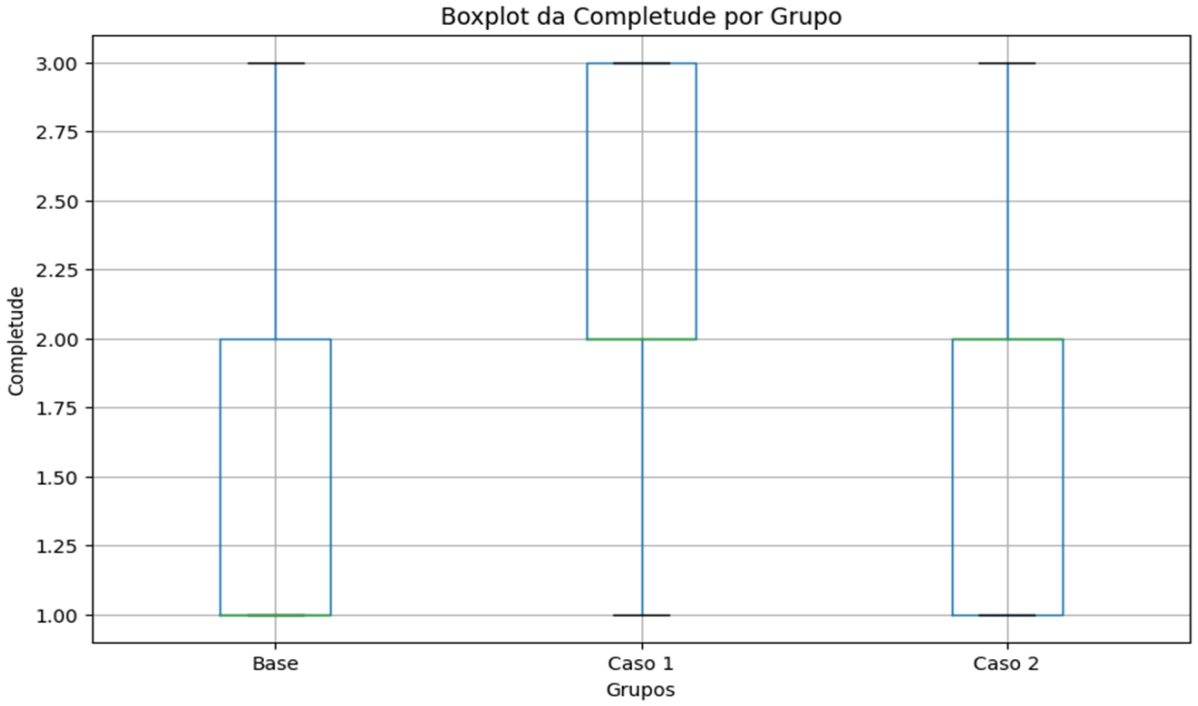
Em relação à completude, o ChatGPT 4.0 sem treinamento (Caso 1) apresentou o maior valor na escala Likert do julgamento das respostas na maioria das questões, seguido pelo ChatGPT 4.0 com treinamento (Caso 2), enquanto a correção do professor (Base) teve a menor completude. Esses resultados sugerem que, embora a IA sem treinamento prévio seja eficiente no julgamento da precisão e completude das respostas, a versão treinada não proporciona melhorias significativas e, em muitos casos, piora o desempenho.

FIGURA 2 – Boxplot da acurácia na correção das questões discursivas pelos diferentes grupos de corretores.



Fonte: Autores.

FIGURA 3 – Boxplot da completude na correção das questões discursivas pelos diferentes grupos de corretores.



Fonte: Autores.

Os resultados encontrados indicam que a versão do ChatGPT 4.0 sem treinamento prévio muitas vezes proporciona uma correção de acurácia comparável à de um professor, enquanto o treinamento adicional parece degradar essa performance. Alguns possíveis *insights* que podem explicar esses achados são:

- Qualidade do treinamento adicional: O treinamento adicional aplicado ao ChatGPT pode não ter sido adequado ou bem calibrado para a tarefa específica de correção de provas discursivas. O treinamento incluiu apenas um exemplo de como a questão deveria ser respondida. Se este treinamento não estiver alinhado com o padrão do professor, isso pode ter introduzido inconsistências e erros adicionais.
- Sobrecarga de informações: O treinamento adicional pode ter introduzido uma sobrecarga de informações ou regras complexas que dificultam a capacidade da IA de tomar decisões claras e precisas. Isso pode levar a uma diminuição na acurácia, já que o modelo pode tentar aplicar padrões inadequados ou contraditórios.
- Adaptabilidade da IA sem treinamento: O ChatGPT 4.0 sem treinamento adicional, ao utilizar um modelo mais geral, pode ser mais adaptável e flexível em suas respostas, resultando em uma acurácia mais alta. Isso sugere que a base de conhecimento e os algoritmos padrão do ChatGPT são já bastante robustos e eficientes para a tarefa de correção de provas discursivas

Esses *insights* destacam a importância de avaliar cuidadosamente os métodos e conteúdos utilizados no treinamento adicional de IA garantindo que eles realmente aprimorem a performance em tarefas específicas.

Em relação às limitações deste estudo, é importante destacar que os dados utilizados como referência, ou seja, as correções realizadas pelo educador, foram conduzidos por um docente habilitado na disciplina em questão. Para futuras pesquisas, recomendamos a ampliação do número de fontes de dados utilizados como base, o que pode enriquecer as comparações e proporcionar uma visão mais abrangente sobre a eficácia das correções automatizadas.

5. CONCLUSÃO

A incorporação de tecnologias de Inteligência Artificial (IA) na educação, especialmente nas universidades, apresenta um potencial significativo para transformar os métodos de avaliação. Este estudo investigou a eficácia do ChatGPT 4.0 na correção de questões discursivas, tanto sem quanto com treinamento prévio.

Os resultados indicam que a versão do ChatGPT 4.0 sem treinamento prévio frequentemente oferece uma correção de acurácia comparável à realizada por professores humanos, sem diferenças significativas na maioria das questões. Em contraste, o treinamento adicional não apenas falhou em melhorar a acurácia, mas frequentemente a degrada. Essa descoberta sugere que o modelo padrão do ChatGPT 4.0 já é robusto e eficaz para a tarefa de correção de provas discursivas, enquanto o treinamento adicional pode introduzir complexidades desnecessárias ou inconsistências.

Em relação à completude, o ChatGPT 4.0 sem treinamento apresentou as maiores pontuações no julgamento da maioria das questões, seguido pela versão treinada, e por fim, pela correção dos professores. Este padrão pode ser atribuído à adaptabilidade e flexibilidade do modelo padrão da IA, que é capaz de fornecer avaliar respostas mais abrangentes sem o viés introduzido pelo treinamento específico.

Os *insights* obtidos ressaltam a necessidade de uma abordagem crítica e cuidadosa na implementação de treinamentos adicionais para modelos de IA em contextos educacionais. A qualidade e a calibração do treinamento são cruciais para garantir que a IA realmente melhore sua performance e não comprometa a precisão das correções.

Em conclusão, o uso de IA generativa como o ChatGPT 4.0 na avaliação educacional demonstra um grande potencial, desde que utilizada de forma adequada. A implementação de tais ferramentas devem ser acompanhadas por um entendimento profundo de suas capacidades e

limitações, garantindo que a inovação tecnológica contribua positivamente para o ambiente educacional. Este estudo contribui para a discussão sobre a integração de IA na educação, destacando a importância de estratégias de adoção bem-informadas e baseadas em evidências.

REFERÊNCIAS

Aleksandr Tiulkanov, AI and Data Policy Lawyer, enero de 2023.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 267–270).

Baker, R. S., Berning, A. W., Gowda, S. M., Zhang, S., & Hawn, A. (2020). Predicting K-12 dropout. Journal of Education for Students Placed at Risk (JESPAR), 25(1), 28–54.

CHEW, E.; CHUA, X.N. (2020). Experiências com IA e o Relatório Horizon. Appl. Sci., v. 11, n. 11, p. 5467.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. In Machine Learning and the City: Applications in Architecture and Urban Design (pp. 535–545).

Janpla, S.; Piriyasurawong, P. (2020). "Titulo do Estudo". Revista Científica, Volume, Número, Página.

Kizilcec, R. F. (2023). To Advance AI Use in Education, Focus on Understanding Educators. International Journal of Artificial Intelligence in Education, 34(1), 12-19. doi: 10.1007/s40593-023-00351-4.

Mirchi, N.; Bissonnette, V.; Yilmaz, R.; Ledwos, N.; Winkler-Schwartz, A.; Del Maestro, R.F. (2020). "Titulo do Estudo". Revista Científica, Volume, Número, Página.

Ribera, M., & Díaz Montesdeoca, O. (Coords.). (2024). ChatGPT y educación universitaria: posibilidades y límites de ChatGPT como herramienta docente. Colección Educación Universitaria. Barcelona: Ediciones Octaedro.

Rodríguez-Ascaso, A.; Boticario, J.G.; Finat, C.; Petrie, H. (2017). "Titulo do Estudo". Revista Científica, Volume, Número, Página.

Samarakou, M.; Fylladitakis, E.D.; Karolidis, D.; Früh, W.-G.; Hatziapostolou, A.; Athinaios, S.S.; Grigoriadou, M. (2016). "Titulo do Estudo". Revista Científica, Volume, Número, Página.

Sabzalieva, E. y Valentini, A. (2023). ChatGPT e inteligencia artificial en la educación superior: guia de início rápido. https://unesdoc.unesco.org/ark:/48223/pf0000385146_spa

Saplacan, D.; Herstad, J.; Pajalic, Z. (2018). "Titulo do Estudo". Revista Científica, Volume,

Número, Página.

UNESCO. (2020). Relatório sobre IA na Educação. Paris.

.