Avaliação de Correções de Inteligência Artificial Generativa no Processo de Escrita de Redações aplicadas ao Exame Nacional do Ensino Médio (ENEM)

Adrielly Mirella Paixão, Marcelo Damasceno de Melo

Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (IFRN)
São Gonçalo do Amarante – RN – Brasil

paixaoadrielly@gmail.com, marcelo.damasceno@ifrn.edu.br

Abstract. *This paper evaluates Generative AIs (ChatGPT and Gemini) in grading essays based on the scoring model of the Exame Nacional do Ensino Médio (ENEM). The methodology involved collecting 338 essays from UOL, which were evaluated according to ENEM's scoring criteria. After gathering the samples, the models graded the essays based on ENEM's Competencies I and III. The scores assigned by human evaluators were compared to those generated by AI models. The results, based on the QWK and MAE metrics, indicated that Generative AIs do not achieve satisfactory performance and they are prompt-sensitive. However, they can serve as a complementary tool for students without access to human evaluators.*

Resumo. *Este trabalho avalia a eficácia de IAs Generativas (ChatGPT e Gemini) ao corrigir redações, seguindo o modelo de correção do Exame Nacional do Ensino Médio (ENEM). A metodologia consistiu na coleta de 338 redações do UOL que foram avaliadas conforme os critérios do exame. Os modelos pontuaram as redações com base na Competência I e III do ENEM. As notas atribuídas pelo corretor humano foram então comparadas às notas geradas pelo modelo. Os resultados, baseados nas métricas QWK e MAE, indicaram que as IAs Generativas não atingem desempenho satisfatório e que são sensíveis ao prompt utilizado. No entanto, elas podem atuar como ferramenta complementar para estudantes sem acesso a corretores humanos.*

1. Introdução

O Exame Nacional do Ensino Médio (ENEM) representa um dos principais métodos de acesso ao ensino superior no Brasil [Junqueira, Martins and Lacerda 2017]. Entre suas cinco provas, a redação assume um papel central na nota final dos candidatos, podendo elevar significativamente suas chances de ingresso em universidades públicas e privadas. No entanto, pesquisas indicam que o desempenho no ENEM está diretamente relacionado ao nível socioeconômico dos participantes. Estudantes provenientes de famílias com maior renda tendem a obter melhores resultados no exame [Lucena and Santos 2020], e a renda familiar se configura como um fator determinante para o acesso ao ensino superior [Andrade 2012].

Dados disponibilizados pelo INEP, instituição responsável pela organização do ENEM, apontam que, embora 2.308 participantes tenham obtido notas entre 980 e 1.000 na redação do ENEM, apenas 215 eram oriundos da rede pública. Além disso, no

intervalo de 950 a 980 pontos, apenas 4.483 estudantes de escolas públicas alcançaram essa pontuação, contra 27.430 de escolas particulares [Secretaria da Comunicação Social 2025]. Essa discrepância evidencia a influência de fatores como a limitação de recursos didáticos, a falta de acompanhamento especializado e a ausência de oportunidades para receber correções detalhadas. Nesse contexto, a Inteligência Artificial (IA) tem sido explorada como uma ferramenta potencialmente capaz de democratizar o acesso a correções de redação, oferecendo avaliações automatizadas e sugestões para aprimoramento textual.

No campo da educação, o uso de IA em assistentes virtuais tornou-se popular, sendo o ChatGPT um dos exemplos mais notáveis, utilizado para diversas finalidades, incluindo pesquisa, programação e assistência na escrita de textos [Jalil et al. 2023]. Levando em consideração essas possibilidades, este estudo propôs avaliar a eficácia do ChatGPT e de outras IAs generativas como assistentes na correção de redações do ENEM.

Este estudo tem como objetivo principal entender o potencial e as limitações dessas ferramentas no processo de aprendizagem da escrita, através da comparação das notas atribuídas pelas IAs com as notas dos corretores humanos. Além disso, o estudo busca contribuir com a discussão sobre o impacto da IA na educação e sua possível aplicação como um recurso acessível para estudantes que não possuem acesso a corretores humanos.

2. Inteligência artificial e educação

Segundo [Jovanović and Campbell 2022], a modelagem generativa é uma técnica de IA que analisa dados de treinamento para aprender padrões e distribuições, permitindo a criação de novos conteúdos realistas. A IA generativa (GAI) utiliza essa abordagem aliada ao aprendizado profundo para produzir textos, imagens, áudio e vídeos em larga escala. De acordo com [Baidoo-Anu and Ansah 2023], modelos generativos possuem duas maiores técnicas: a Generative Adversarial Network (GAN) e a Generative Pre-trained Transformer (GPT). A GAN possui duas redes neurais: uma geradora, que cria conteúdos e outra discriminadora, que julga se o conteúdo gerado é real ou não, e este processo acontece até que a rede discriminadora não saiba distinguir o conteúdo artificial do real. Já o GPT cria novos conteúdos textuais a partir de artigos digitais existentes.

Na educação superior, a IA tem sido discutida tanto como ferramenta de apoio quanto como possível ameaça à integridade acadêmica. Alguns estudos sugerem que tecnologias como o ChatGPT podem transformar as formas tradicionais de ensino e aprendizagem [Rudolph, Tan and Tan 2023]. Enquanto outros, enfatizam os riscos de plágio e desvalorização do processo de aprendizagem pelos alunos. Uma vez que um aluno pede para uma IA faça sua atividade acadêmica, este perde a chance de aprender [Cotton, Cotton and Shipway 2023]. Porém, de acordo com [Nazari, Shabbir and Setiawan 2021], assistentes digitais equipados com IA não apenas auxiliam estudantes na escrita acadêmica, mas também aumentam sua confiança e engajamento na escrita,

por conta feedbacks formativos e instrucionais em tempo real, além da revisão detalhada de seus textos.

No contexto da avaliação automática de textos, [Mizumoto and Eguchi 2023] definem *Automated Essay Scoring* (AES) ao uso de algoritmos para avaliar e pontuar redações com base em critérios predefinidos, como correção linguística, riqueza lexical, coerência, sintaxe e relevância semântica. Os autores concluem que o uso do AES em conjunto ao GPT tem o potencial de revolucionar o ensino e a avaliação de redações, fornecendo pontuação e *feedback* imediato, proporcionando redução na carga de trabalho dos professores, o que permite que se concentrem em outras tarefas.

Por fim, a IA na educação apresenta benefícios e riscos, dependendo de como é utilizada. Enquanto tecnologias como o ChatGPT podem impulsionar a aprendizagem por oferecer tutoria personalizada que está sempre à disposição do aluno, também levantam preocupações, como a dependência em dados desses modelos [Baidoo-Anu and Ansah 2023]. Assim, o impacto da IA na educação dependerá diretamente de como seus recursos são integrados de forma consciente e alinhada aos objetivos pedagógicos [Harry 2023].

3. Critérios de correção do ENEM

A correção da redação do ENEM é baseada em cinco competências [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) 2024]:

I. demonstrar domínio da modalidade escrita formal da língua portuguesa;
II. compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema dentro dos limites estruturais do texto dissertativo-argumentativo em prosa;
III. selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista;
IV. demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação;
V. elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

Cada competência é avaliada com notas de 0 a 200 pontos, em intervalos de 40 pontos. Dois avaliadores atribuem notas para cada competência, e a nota final é a média das notas mais próximas. Caso haja discrepância de 80 pontos em qualquer competência ou 100 pontos na nota final, um terceiro avaliador é acionado. Persistindo a divergência, a redação é encaminhada a uma banca especial. Um corretor atribui nota zero à redação se: (1) o candidato fuja totalmente do tema proposto; (2) escreva um texto que não seja do tipo dissertativo-argumentativo; ou (3) produza menos de 7 linhas de texto.

Neste estudo, optou-se por focar nas Competências I e III. A Competência I avalia o domínio da escrita formal da língua portuguesa e a fluidez do texto, enquanto a Competência III examina a capacidade de defender um ponto de vista por meio da seleção, organização e interpretação de informações, fatos, opiniões e argumentos. A escolha dessas competências se justifica por representarem, respectivamente, aspectos

estruturais e argumentativos fundamentais da redação dissertativo-argumentativa exigida no ENEM.

4. Metodologia

Esta seção descreve os procedimentos adotados para a coleta, o processamento e a avaliação dos textos gerados, bem como as estratégias para comparar as avaliações da IA com as dos revisores humanos.

4.1 Coleta de dados

A coleta de dados consistiu na extração de redações do Banco de Redações do UOL por meio de técnicas de *web scraping*. Para isso, foram utilizadas as bibliotecas Requests e BeautifulSoup4, ambas na linguagem Python. BeautifulSoup4 (bs4) é uma biblioteca que facilita a extração de dados de um arquivo HTML ou XML. Com a biblioteca Requests, foi realizada uma requisição HTTP para obter o código-fonte da página. Em seguida, a resposta foi processada com a bs4, identificando as *tags* e classes que continham os textos das redações, as notas das Competências 1 e 3, o título da redação e o título do tema. A Figura 1 apresenta a estrutura HTML identificada no site para a extração do título e texto da redação.

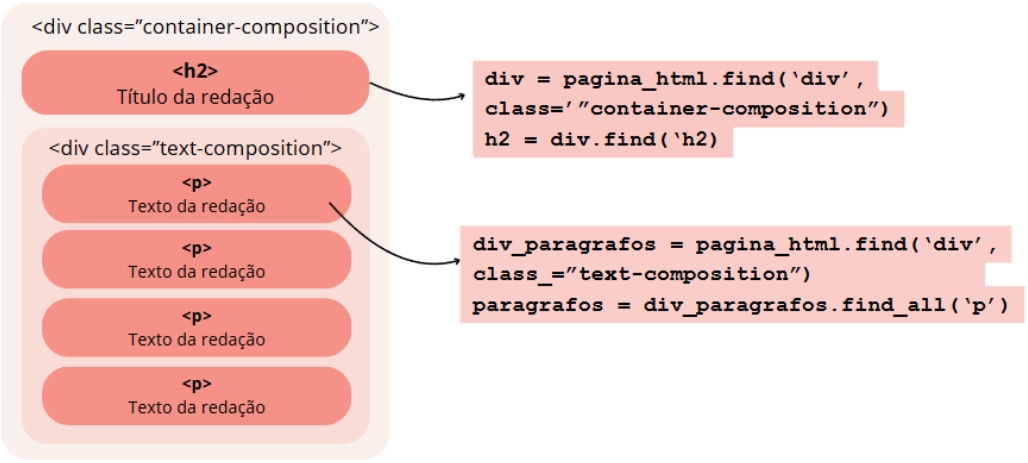

Figura 1. Estrutura HTML e código Python para extração de título e parágrafos com BeautifulSoup4.

Por fim, os textos foram extraídos e armazenados em arquivos de texto para posterior análise. Para evitar bloqueios por parte do servidor durante a coleta de dados, foi inserido um intervalo de tempo entre as requisições. Este tempo foi randomizado, tornando o padrão de acesso ao site menos previsível e mais próximo do comportamento humano.

Inicialmente, estimava-se coletar 1.120 redações de 56 temas disponíveis, mas o conjunto final ficou em 338 textos distribuídos em 17 temas distintos. Essa redução ocorreu devido a dois fatores principais: (1) 12 dos 56 temas disponíveis não possuíam redações cadastradas e (2) diferenças entre os critérios de pontuação da UOL e do ENEM exigiram a exclusão de 27 temas. No ENEM, cada competência é pontuada em

intervalos de 40 pontos, enquanto, nas redações dos 27 temas, as competências foram avaliadas em intervalos de 50 pontos. Por não seguir os critérios do ENEM, optou-se por excluir tais redações. Apesar dessa redução, a amostra final permitiu uma avaliação representativa das capacidades dos modelos analisados.

4.2 Correção das redações pelos modelos

Nesta pesquisa, foram escolhidos os modelos ChatGPT (GPT-4o-mini) e Gemini (gemini-1.5-pro). Optou-se pelo uso dos modelos devido à sua acessibilidade e viabilidade econômica. O gemini-1.5-pro foi selecionado por estar disponível gratuitamente via API, enquanto o GPT-4o-mini se destacou por oferecer desempenho a um custo reduzido, tornando ambos adequados para experimentos nos requisitos deste trabalho. As APIs possuem diferentes políticas de acesso: a API do Gemini, limitada a 2 requisições por minuto e 50 por dia, exigiu controle de taxa; já a API da OpenAI adota cobrança por tokens, sem restrições de frequência. Para cada redação, foram realizadas duas requisições (Competência I e III), cujas respostas foram consolidadas em um único arquivo TXT.

A estrutura do prompt utilizado no pedido de correção das redações é mostrada na Figura 2, e foi organizada de modo a conter: o objetivo da solicitação, especificando a tarefa a ser realizada pelo modelo; estrutura que o modelo deve retornar a resposta; alertas que indicam aspectos críticos a serem observados; informações de contexto com detalhes adicionais para aprimorar a precisão da resposta gerada.

```
Lendo a redação abaixo, eu quero que você dê uma nota para a Competência [I ou
III], seguindo o modelo de correção do ENEM.

[Redação]

Corrija a redação seguindo o seguinte modelo:
"Nota para Competência [número da competência]: [valor]
Justificativa para a nota: [...]". A correção deve conter apenas esses campos.

CUIDADO: Tenha certeza que a redação não foge do tema proposto. Considera-se
que uma redação tenha fugido ao tema quando nem o assunto mais amplo nem o tema
específico proposto tenham sido desenvolvidos. Caso a redação fuja do tema,
atribua a ela a nota zero.

Contexto: Para dar nota para a Competência, use como base o texto a seguir.
[Texto explicando a Competência I ou III]
```

Figura 2. Estrutura do *prompt* utilizado

Adicionalmente, foi também utilizado outro modelo de *prompt*, sem a seção de "CUIDADO" que pede que a redação seja atribuída nota zero caso fuja do tema proposto, como mostra a Figura 3.

```
Lendo a redação abaixo, eu quero que você dê uma nota para a Competência [I ou
III], seguindo o modelo de correção do ENEM.

[Redação]

Corrija a redação seguindo o seguinte modelo:
"Nota para Competência [número da competência]: [valor]
Justificativa para a nota: [...]". A correção deve conter apenas esses campos.

Contexto: Para dar nota para a Competência, use como base o texto a seguir.
[Texto explicando a Competência I ou III]
```

Figura 3. Estrutura do *prompt* sem critério de fuga total ao tema

Durante o processo de coleta das respostas geradas pelo modelo, observou-se que o Gemini atribuía notas zero às redações com maior frequência, enquanto o ChatGPT raramente atribuía essa pontuação. Para a pesquisa, foi considerada apenas a possibilidade da fuga ao tema. A fuga total do tema consiste em escrever sobre um assunto que não corresponde ao tema proposto na prova. Com isso, o texto que indica que a redação deve ser zerada caso haja fuga do tema foram removidos do solicitação, e novas requisições foram realizadas com essa versão modificada do prompt.

4.3 Avaliação dos modelos

Para avaliar a eficácia dos modelos generativos na correção de redações do ENEM, este estudo baseou-se na abordagem de [da Silva Júnior 2021], que buscou criar um modelo que automatiza a avaliação de redações, com foco nas Competências I e III do ENEM. O estudo desenvolveu um modelo baseado em Regressão Linear combinada com Algoritmos Genéticos (RL+GA), buscando reproduzir avaliações semelhantes a corretores humanos. Os resultados foram analisados com base nas mesmas métricas utilizadas neste trabalho: Kappa Quadrático Ponderado (ou QWK – Q*uadratic Weighted Kappa*) e Erro Médio Absoluto (ou MAE - *Mean Absolute Error*). Estas métricas foram adotadas neste trabalho para comparar o desempenho do ChatGPT e do Gemini na tarefa de atribuição de notas pois permitem avaliar tanto a proximidade das notas quanto o nível de concordância entre avaliadores,

O Mean Absolute Error (MAE) é uma métrica amplamente utilizada na avaliação de modelos de regressão [Hodson 2022]. Envolve somar os valores absolutos dos erros para obter o 'erro total' e, em seguida, dividir esse erro total por n [Wilmott and Matsuura 2005]. Para uma amostra de n observações y (yi, com $i = 1, 2, ..., n$) e n previsões correspondentes do modelo $\widehat{y}$, o MAE é [Hodson 2022]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y}_i| .$$

Para o estudo, as observações y foram as notas dadas pelos corretores humanos, e as previsões $\widehat{y}$, as notas dadas pelo modelo.

O QWK é uma métrica estatística utilizada para quantificar o grau de concordância entre dois avaliadores, ajustando a concórdia esperada ao acaso e ponderando a gravidade dos desacordos [Li, Gao and Yu 2023]. Essa medida é especialmente relevante em cenários onde os escores são ordinais, como no contexto de correção automática de redações (Automated Essay Scoring, ou AES) [Doewes, Kurdhi and Saxena 2023].

O QWK difere das versões simples ou linearmente ponderadas do kappa ao atribuir pesos quadráticos à discordância entre as categorias. Ou seja, quanto maior a distância entre os escores atribuídos por dois avaliadores, maior a penalidade na pontuação do QWK [Li, Gao and Yu 2023] . Formalmente, o peso quadrático entre duas categorias i e j é definido como:

$$w_{ij} = 1 - \left(\frac{i-j}{n-i}\right)^2,$$

onde n é o número de categorias possíveis. O QWK é então calculado como:

$$QWK = 1 - \frac{\Sigma_{ij} w_{ij} O_{ij}}{\Sigma_{ij} w_{ij} E_{ij}},$$

em que O_{ij} representa a matriz de observações reais entre os dois avaliadores, e E_{ij} a matriz de concordância esperada ao acaso, baseada na distribuição marginal dos escores.

As notas atribuídas por cada modelo generativo foram comparadas individualmente às notas do corretor humano, utilizando as métricas MAE e QWK. As métricas foram implementadas com a biblioteca scikit-learn, uma das principais ferramentas de aprendizado de máquina em Python, que também oferece ferramentas para pré-processamento e análise de dados.

5. Resultados e discussões

Neste estudo, a análise dos modelos ChatGPT e Gemini revelou que a capacidade desses modelos de avaliar redações do ENEM é limitada. Os resultados obtidos para as métricas QWK e MAE indicam baixa concordância entre as notas das IAs e as atribuídas por corretores humanos, como é observado nas Tabelas 1 e 2.

Tabela 1. Resultado das métricas QWK e MAE para a Competência I

Modelo	QWK	MAE
ChatGPT	-0.00464	47.8698
ChatGPT (sem avisos de fuga do tema)	0.0	47.8698
Gemini	0.0297	88.1065
Gemini (sem avisos de fuga do tema)	0.3277	38.2840

Conforme é mostrado na Tabela 1, os valores de QWK para a Competência I sem a seção de avisos no prompt foram de -0.00464 para o ChatGPT e 0.0297 para o Gemini, sugerindo que as notas atribuídas estão praticamente aleatórias em relação à avaliação humana. Com a remoção do critério de fuga ao tema no prompt, houve uma melhoria no Gemini (QWK = 0.3277), enquanto o ChatGPT manteve resultados semelhantes. Ademais, a retirada do critério de fuga ao tema reduziu consideravelmente o MAE do Gemini, que passou de 88.1065 para 38.2840. Isso indica que o modelo apresentou menor discrepância em relação às notas humanas quando essa instrução não foi incluída no prompt. Esse resultado sugere que a forma como os critérios são apresentados à IA pode influenciar diretamente sua capacidade de pontuar as redações com maior precisão.

Tabela 2 - Resultado das métricas QWK e MAE para a Competência III

Modelo	QWK	MAE
ChatGPT	0.03927	42.8402
ChatGPT (sem avisos de fuga do tema)	0.02862	43.4319
Gemini	0.01009	99.1715
Gemini (sem avisos de fuga do tema)	0.38591	44.7337

Na tabela 2, pode-se observar que os valores QWK do ChatGPT para a Competência III continuam baixos, independentemente da presença de instrução de fuga do tema no prompt, indicando novamente a aleatoriedade do modelo para a avaliação das redações. O MAE do ChatGPT também teve pouca alteração, o que sugere que a modificação no prompt não influenciou significativamente sua capacidade de prever notas mais próximas das atribuídas por avaliadores humanos. Já o Gemini, nota-se que os valores de QWK mudaram expressivamente com a remoção do critério de fuga ao tema. Esse aumento indica que o modelo tornou-se mais consistente na pontuação das redações quando não havia uma orientação explícita sobre fuga ao tema. Da mesma forma, o MAE do Gemini também reduziu, o que demonstra uma melhora na precisão das notas atribuídas. Esses resultados sugerem que o Gemini é mais sensível à formulação dos prompts do que o ChatGPT, apresentando variações significativas dependendo das instruções fornecidas.

Os valores de QWK obtidos pelos modelos generativos neste estudo foram significativamente inferiores aos reportados por [da Silva Júnior 2021], que aplicou um modelo baseado em Regressão Linear e Algoritmos Genéticos para a correção de redações. Enquanto seu modelo alcançou um QWK de 0.7484 para a Competência I e 0.619 para a Competência III, os valores observados no ChatGPT e no Gemini foram muito menores, indicando uma baixa concordância com as notas humanas. Esse resultado sugere que modelos treinados especificamente para a tarefa de avaliação de redações podem ser mais eficazes do que modelos generativos generalistas.

Os resultados obtidos demonstram que os modelos generativos avaliados não são exclusivamente adequados para a correção automatizada de redações do ENEM na configuração testada. A análise sugere que a IA generativa, por ser treinada com um grande volume de dados diversos, não possui especialização suficiente para pontuar redações com base nos critérios específicos do ENEM. Ainda não sabemos qual seria a melhor abordagem para tornar modelos generativos mais eficazes na avaliação de redações. Sendo necessário utilizar outros modelos, ferramentas ou metodologia distinta. No entanto, sugerimos que essas tecnologias sejam utilizadas de forma confiável, seria necessário um modelo treinado especificamente para a correção de redações do ENEM e métodos que garantam maior estabilidade e precisão na atribuição de notas. Requisitos estes que o ChatGPT e o Gemini não atendem. No entanto, o seu uso não pode ser descartado pelos estudantes, pois são ferramentas úteis de apoio e a escrita de redações.

6. Considerações finais

Este estudo buscou investigar a eficácia de IAs generativas na correção de redações do ENEM, analisando a precisão das notas atribuídas pelos modelos em relação às notas de avaliadores humanos. Os resultados indicam que, com os prompts utilizados, os modelos não são indicados para correções, apresentando baixo índice de concordância (QWK) e valores consideráveis de erro (MAE). Acredita-se que o principal obstáculo está na falta de treinamento específico dos modelos para o contexto do ENEM. Além disso, melhorias na formulação dos prompts podem contribuir para aumentar a precisão das respostas geradas pelos modelos. Os resultados mostram que pequenas modificações nos prompts podem afetar significativamente a avaliação das redações, sugerindo que a estruturação das instruções é um fator importante na precisão dos modelos.

A pesquisa busca contribuir no debate do uso de Inteligências Artificiais no ensino e o seu potencial de promover o processo de aprendizagem da escrita. Futuras pesquisas podem explorar a adaptação dos modelos para esse propósito, criando um modelo próprio para correções, treinada com um conjunto de dados selecionado a fim de obter melhores resultados.

Referências

Andrade, C. Y. (2012). Acesso ao ensino superior no Brasil: equidade e desigualdade social. *Revista Ensino Superior Unicamp*, 6, 18-27.

Baidoo-Anu, D. and Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52-62.

INEP. (2024). A redação no Enem 2024: Cartilha do participante. Brasília, DF: INEP. https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/a_redacao_no_enem_2024_cartilha_do_participante.pdf

Cotton, D. R., Cotton, P. A., and Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in education and teaching international*, *61*(2), 228-239.

da Silva Júnior, J. A. (2021). Um avaliador automático de redações. *Dissertação de Mestrado, Universidade Federal do Espírito Santo*.

Doewes, A., Kurdhi, N., & Saxena, A. (2023, July). Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *16th International Conference on Educational Data Mining, EDM 2023* (pp. 103-113). International Educational Data Mining Society (IEDMS).

Harry, A. (2023). Role of AI in Education. *Interdiciplinary Journal & Hummanity (INJURITY),* 2(3).

Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, 2022, 1-10.

Jalil, S. et al. (2023). Chatgpt and software testing education: Promises & perils. In: *2023 IEEE international conference on software testing, verification and validation workshops (ICSTW)*, pages 4130-4137. IEEE.

Jovanović, M. and Campbell, M. (2022). Generative artificial intelligence: Trends and prospects. In *Computer*, pages 107-112, vol. 55.

Junqueira, R. D., Martins, D. A., & Lacerda, C. B. F. (2017). Política de acessibilidade e exame nacional do ensino médio (ENEM). *Educação & Sociedade*, 38, 453-471.

Li, M., Gao, Q., & Yu, T. (2023). Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. *BMC cancer*, *23*(1), 799.

Lucena, J. P. O. and Santos, H. N. L. (2020). A relação entre desempenho no Exame Nacional do Ensino Médio e o perfil socioeconômico: um estudo com os microdados de 2016. *Revista de Gestão e Secretariado*, *11*(2), 1–23.

Mizumoto, A. and Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050.

Nazari, N., Shabbir, M. S. and Setiawan, R. (2021). Application of Artificial Intelligence powered digital writing assistant in higher education: randomized controlled trial. *Heliyon*, *7*(5).

Rudolph, J., Tan, S. and Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of applied learning and teaching*, *6*(1), 342-363.

Secretaria de Comunicação Social. (2025). *Enem 2024: Resultados mostram crescimento na adesão e na média das notas*. Governo do Brasil. https://www.gov.br/secom/pt-br/assuntos/noticias/2025/janeiro/enem-2024-resultados-mostram-crescimento-na-adesao-e-na-media-das-notas.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79-82.