Estudo de Caso

ISSN 2177-8310
DOI: 10.18264/ eadf.v16i1.2556

EaD em foco
Revista Científica em Educação a Distância

Análise Comparativa de Ferramentas de IA Generativa na Correção de Redações Nota 1000 no ENEM

Comparative Analysis of Generative AI Tools in the Correction of Essays with a Score of 1000 in ENEM

Jorge Luis Cavalcanti RAMOS[1]*
Hebert Henrique Barboza de BRITO[1]
João Carlos Sedraz SILVA[1]
André Gonçalves MARTINS[1]
Rodrigo Lins RODRIGUES[2]

[1]Universidade Federal do Vale do São Francisco – Juazeiro-BA – Brasil.
[2]Universidade Federal Rural de Pernambuco – Recife-PE – Brasil
*jorge.cavalcanti@univasf.edu.br

Resumo

Este estudo apresenta a análise comparativa de quatro plataformas de Inteligência Artificial (IA) generativa, na correção de 100 (cem) redações no ENEM, para as quais foram atribuídas notas 1000 (mil) pelos avaliadores oficiais do INEP. O objetivo foi avaliar quatro diferentes plataformas para a correção de redações com nota máxima e comparar essas avaliações em relação à avalição oficial, que atribuiu nota máxima em todas. O percurso metodológico incluiu a criação de uma base de dados com os temas e textos das cem redações, a criação de scripts para interação com as plataformas e o armazenamento dos resultados com os feedbacks de cada redação, a partir dos critérios oficiais, para posterior análise dos resultados. Para as análises, foram usadas estatísticas descritivas e estatísticas inferenciais por meio dos testes ANOVA e Tukey. Os resultados apontaram a plataforma brasileira Maritaca IA como a que obteve os melhores resultados, com as notas mais próximas às atribuídas pelos avaliadores do ENEM, sugerindo seu uso em aplicações na qual o idioma português e o contexto brasileiro devem ser considerados.

Palavras-chave: Redação. ChatGPT. Gemini. DeepSeek. Maritaca. Correção Automática.

CC BY

Recebido: 25/06/2025
Aceito: 13/01/2026
Publicado: 26/03/2026
Editores Responsáveis: Daniel Salvador/ Carmelita Portela/ Daniela Samira

COMO CITAR ESTE TRABALHO

RAMOS, J. L. C. *et al.* Análise Comparativa de Ferramentas de IA Generativa na Correção de Redações Nota 1000 no ENEM. **EaD Em Foco**, 16(1), e2556. Doi: https://doi.org/10.18264/eadf.v16i1.2556

Comparative Analysis of Generative AI Tools in the Correction of Essays with a Score of 1000 in ENEM

Abstract

This study presents a comparative analysis of four generative Artificial Intelligence (AI) platforms in the correction of 100 essays in the ENEM, which were awarded a score of 1000 by official INEP evaluators. The objective was to evaluate four different platforms for correcting essays with maximum scores and compare these evaluations with the official evaluation, which assigned maximum scores to all of them. The methodological approach included the creation of a database with the themes and texts of the one hundred essays, the creation of scripts for interaction with the platforms, and the storage of the results with feedback for each essay, based on official criteria, for subsequent analysis of the results. Descriptive statistics and inferential statistics were used for the analyses, using ANOVA and Tukey tests. The results pointed to the Brazilian platform Maritaca IA as the one that obtained the best results, with scores closest to those assigned by ENEM evaluators, suggesting its use in applications in which the Portuguese language and Brazilian context must be considered.

Keywords: *Essay. ChatGPT. Gemini. DeepSeek. Maritaca. Automatic correction.*

1. Introdução

Ingressar em uma universidade pública é um objetivo almejado por muitos jovens brasileiros, sendo a obtenção de uma boa nota no Exame Nacional do Ensino Médio (ENEM) um dos principais desafios para alcançar esse propósito. Nesse contexto, a redação assume papel central, pois é avaliada a partir de critérios específicos que orientam tanto a produção textual dos estudantes quanto o processo de correção realizado por especialistas, os quais são continuamente aprimorados.

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) divulga anualmente redações que obtiveram nota máxima como referência para estudantes e professores, contribuindo para o aprimoramento das estratégias de preparação para esse componente do exame. Paralelamente, o avanço das plataformas de Inteligência Artificial generativa, impulsionado a partir do lançamento do ChatGPT em 2022, ampliou o uso dessas tecnologias como ferramentas de apoio à escrita, correção e otimização de tarefas, com o surgimento de diversas soluções desenvolvidas por grandes empresas e *startups*.

O uso de Inteligência Artificial (IA), na avaliação educacional, tem sido amplamente discutido na literatura contemporânea, especialmente a partir do avanço dos *Large Language Models* (LLM), que demonstram capacidade consistente de análise textual, identificação de padrões linguísticos, coerência argumentativa e aderência a critérios previamente definidos (Kasneci et al., 2023; Sallam, 2023).

No campo da avaliação de produções escritas, estudos recentes apontam que sistemas baseados em IA podem atuar como instrumentos de apoio formativo, oferecendo feedback estruturado, padronizado

e imediato, sem a pretensão de substituir a avaliação humana, mas contribuindo para a redução da subjetividade, aumento da escalabilidade e democratização do acesso à preparação educacional (Peres et al., 2023; Kehoe, 2023).

No contexto específico do Exame Nacional do Ensino Médio (ENEM), cuja redação é avaliada a partir de critérios oficiais bem definidos e publicamente documentados pelo Inep, a aplicação de IA para fins de correção automática precisa ter sua viabilidade técnica e pedagógica analisada a fundo, a partir de modelos capazes de compreender a língua portuguesa em seu uso formal e no contexto sociocultural brasileiro. Importa destacar que este estudo não se propõe a corrigir redações oficiais do ENEM, tampouco a substituir o processo avaliativo conduzido pelo Inep, mas sim a analisar redações já corrigidas e divulgadas como modelos oficiais, que obtiveram nota máxima (1000), com o objetivo de verificar em que medida plataformas de IA generativa conseguem reproduzir padrões de correção próximos aos adotados por avaliadores especialistas.

Dessa forma, a justificativa do estudo reside tanto no avanço científico sobre o uso responsável de IA na avaliação educacional quanto em sua aplicação prática, ao subsidiar o desenvolvimento de uma ferramenta de apoio que possa ser utilizada por estudantes e professores no processo de preparação para o ENEM, permitindo que os candidatos treinem de forma objetiva, sistemática e alinhada aos critérios oficiais, contribuindo para a melhoria do desempenho nesse componente essencial do exame, sem caráter substitutivo da avaliação humana, mas como recurso complementar de aprendizagem.

1.1. Desafios da redação no ENEM

A redação do ENEM constitui um dos componentes centrais para o ingresso no ensino superior no Brasil, exigindo dos estudantes não apenas domínio da norma padrão da língua portuguesa, mas também competências argumentativas, capacidade de compreensão temática e elaboração de propostas de intervenção social. Trata-se de uma prova que demanda articulação coerente de ideias, clareza discursiva e síntese, considerando ainda as restrições de tempo e extensão textual impostas pelo exame (Brasil, 2024).

A diversidade temática das propostas de redação — que frequentemente envolve questões sociais, culturais, políticas e científicas — impõe aos candidatos a necessidade de repertório sociocultural amplo e atualização constante, além da habilidade de interpretar textos motivadores sem incorrer em cópia ou reprodução acrítica. A exigência de originalidade, aliada ao respeito aos direitos humanos e à diversidade sociocultural, reforça o caráter formativo e crítico da avaliação (Costa; Lelis; Martins, 2023).

Do ponto de vista avaliativo, a correção das redações também envolve desafios relevantes. Apesar de os critérios oficiais e de os processos de formação e padronização dos corretores, a avaliação textual mantém um grau inevitável de subjetividade, especialmente diante do elevado volume de redações corrigidas em curto espaço de tempo. Esse cenário torna particularmente relevante a investigação de instrumentos capazes de apoiar processos avaliativos de forma mais uniforme e escalável, sem descaracterizar o papel do julgamento humano (Costa; Lelis; Martins, 2023).

As redações do ENEM são avaliadas a partir de cinco competências, cada uma com pontuação máxima de 200 pontos, totalizando 1000 pontos. Essas competências abrangem **domínio linguístico, compreensão da proposta, organização argumentativa, uso de mecanismos coesivos e elaboração de proposta de intervenção**, conforme sistematizado na Cartilha do Participante do Inep (Brasil, 2024). A clareza e a formalização desses critérios tornam o exame um objeto adequado para análises comparativas envolvendo processos automatizados de correção textual.

1.2. *Large Language Models* (LLM) e IA Generativa

Os *Large Language Models* (LLM) representam um avanço significativo no campo do processamento de linguagem natural, ao serem capazes de analisar e gerar textos com elevado grau de fluência, coerência e adequação contextual, a partir do treinamento em grandes volumes de dados textuais (Kasneci et al.,

2023). Essa capacidade tem ampliado o uso desses modelos em diferentes domínios, incluindo aplicações educacionais que envolvem análise, produção e avaliação de textos.

No contexto educacional, os LLM têm sido explorados como ferramentas de apoio ao ensino e à aprendizagem, sobretudo pela possibilidade de oferecer feedback automatizado, suporte à escrita e à mediação de processos formativos em diferentes níveis de ensino. No entanto, sua adoção levanta questões relevantes quanto aos limites da automação, à confiabilidade das respostas e à necessidade de mediação pedagógica, especialmente em atividades que envolvem avaliação e autoria textual (Kasneci et al., 2023; Sallam, 2023).

A Inteligência Artificial Generativa, enquanto categoria mais ampla, refere-se a sistemas capazes de produzir conteúdos inéditos — como textos, imagens e outros artefatos digitais — a partir de padrões aprendidos nos dados de treinamento. Diferentemente de modelos discriminativos, esses sistemas não apenas classificam informações, mas geram novas instâncias que refletem regularidades estatísticas dos dados de origem (Peres et al., 2023; Epstein et al., 2023). Essa característica amplia seu potencial de uso, ao mesmo tempo em que exige atenção crítica aos aspectos éticos, aos vieses incorporados nos modelos e à adequação de seus resultados aos contextos socioculturais nos quais são aplicados (Castelli; Manzoni, 2022).

No âmbito deste estudo, o interesse pelos LLM e pela IA generativa reside em sua capacidade de usar critérios avaliativos explícitos, como os utilizados na correção da redação, possibilitando análises comparativas sobre o grau de aproximação entre avaliações automatizadas e humanas, sem desconsiderar os limites inerentes a sistemas probabilísticos e não determinísticos.

2. Metodologia

Nessa seção, são apresentadas as diretrizes que nortearam a pesquisa, como o método utilizado, com a descrição sumária de cada etapa e as plataformas de IA Generativas escolhidas para os testes comparativos com as correções humanas das redações do ENEM.

O percurso começou com a identificação da necessidade de se avaliar as diferentes plataformas de IA Generativa que utilizam LLM, para a atividade de correção de 100 (cem) avaliações que obtiveram nota 1000 (mil) nos Exames de 2013 a 2023, que foram disponibilizadas pelo Inep pouco antes do ENEM 2024. A pesquisa buscou: a) verificar se existem diferenças significativas entre a correção feita por especialistas e a correção feita pelas plataformas; e b) verificar qual das plataformas avaliadas apresentou resultados mais próximos da correção humana para as redações no ENEM da amostra analisada.

2.1. Etapas da Pesquisa

A pesquisa foi desenvolvida a partir de a execução das seguintes etapas:

1. Escolha das Ferramentas de IA Generativas para utilização no estudo. Nessa etapa, buscou-se um aprofundamento no conhecimento das principais ferramentas disponíveis no mercado levando-se em conta aspectos como: documentação clara e consistente, disponibilidade e facilidade de uso das respectivas APIs (*Application Programming Interface*), gratuidade no uso para o limite de transações requeridas para a pesquisa. Para esta etapa, foram escolhidas as quatro plataformas descritas na fundamentação teórica deste estudo (ChatGPT, Gemini, DeepSeek e Maritaca IA), todas utilizando versões gratuitas dos seus modelos:

- **GPT-4o-mini (OpenAI).** O ChatGPT (*Chat Generative Pre-trained Transformer*) é um chatbot conversacional desenvolvido pela OpenAI, que gera texto em resposta a um prompt fornecido por humanos. Ele é baseado em grandes modelos de linguagem (LLMs) que aprendem autonomamente com os dados (Peres et al., 2023).

- **Gemini-1.5-Pro (Google DeepMind).** O Gemini, desenvolvido pelo Google DeepMind, representa um avanço significativo no campo de modelos de linguagem de larga escala (LLMs). Ele se destaca por sua

natureza multimodal, o que significa que foi projetado para compreender e gerar conteúdo em diversas modalidades, como texto, imagens, áudio e vídeo (Gemini, 2025).

● **DeepSeek-R1:Free (DeepSeek).** O DeepSeek é um modelo de linguagem de grande escala (LLM) desenvolvido para aplicações conversacionais avançadas, combinando técnicas avançadas em mento de linguagem natural, com arquiteturas profundas de transformadores (DeepSeek, 2025).

● **Maritaca IA**. A Maritaca é uma plataforma de inteligência artificial (IA) desenvolvida pela empresa brasileira Maritaca AI, especializada em processar e gerar linguagem natural de forma fluida e coerente. Utilizando técnicas avançadas de aprendizado de máquina e redes neurais profundas, a Maritaca é capaz de entender o contexto de conversas, responder perguntas de maneira relevante, com destaque à sua ampla base de treinamento com dados em língua portuguesa (Maritaca, 2025).

A opção pelo uso das versões gratuitas das plataformas de IA generativa analisadas esteve diretamente relacionada às condições de realização da pesquisa. Trata-se de um estudo sem financiamento externo, cujo lócus de investigação está associado ao contexto de uma escola pública, o que impôs restrições orçamentárias e orientou a adoção de ferramentas que garantissem a infraestrutura tecnológica mínima necessária à execução do experimento.

Embora seja reconhecido que versões pagas ou profissionais possam oferecer modelos mais recentes, maior capacidade de processamento ou parâmetros adicionais de ajuste, optou-se pelas versões gratuitas por refletirem de forma mais realista as condições de acesso predominantes no contexto educacional público brasileiro.

Esse fator constitui uma limitação do estudo e pode ter impactado os resultados obtidos; contudo, tal escolha reforça a relevância prática da pesquisa, ao analisar o desempenho de tecnologias efetivamente acessíveis a estudantes e instituições públicas de ensino.

2. Criação de banco de dados de redação nota 1000. Foram coletados e armazenados em um banco de dados MySQL, os dados das 100 redações disponibilizadas, contendo informações como tema, autor, ano e o texto da redação. As redações foram realizadas no período de 2014 a 2023.

3. Criação de uma aplicação de suporte, para comunicação com as plataformas por meio de suas respectivas APIs. Uma aplicação *web* foi desenvolvida usando a linguagem Python, para permitir a interação com as plataformas e realizar o processamento das correções das redações, gerando arquivos de saída com os resultados.

Com o objetivo de garantir a comparabilidade dos resultados e a replicabilidade do estudo, foi adotado um *prompt* único e padronizado para todas as plataformas de IA generativa analisadas. O conteúdo semântico do *prompt* foi mantido rigorosamente idêntico entre as quatro plataformas, variando apenas os aspectos técnicos necessários à comunicação com cada API, tais como estrutura de mensagens, parâmetros de requisição e sintaxe específica exigida por cada fornecedor. Não foram utilizados ajustes finos, instruções adicionais ou parâmetros diferenciados que pudessem influenciar o comportamento dos modelos de forma desigual entre as plataformas.

O texto base do *prompt* utilizado em todas as interações foi o seguinte: "*Você é um sistema especializado na avaliação de redações do Exame Nacional do Ensino Médio (ENEM). Avalie a redação fornecida com base nos cinco critérios abaixo. Para cada critério, forneça uma nota de 0 a 200. No final, calcule a nota total (soma das notas dos cinco critérios).*" Em seguida, eram explicitados no *prompt* os cinco critérios oficiais de correção do ENEM, conforme descritos na Cartilha do Participante do Inep, sem qualquer adaptação conceitual.

Essa padronização assegurou que as diferenças observadas nos resultados decorrem do comportamento dos modelos e de suas bases de treinamento, e não de variações no comando de entrada, reforçando a validade metodológica e a possibilidade de reprodução do experimento por outros pesquisadores.

4. Entrada, processamento e saída de dados. A aplicação lê um arquivo tipo JSON na base de dados, contendo as informações das diversas redações armazenadas. O usuário (pesquisador) podia escolher qual das plataformas vai utilizar para corrigir as redações. Em seguida, a plataforma faz a correção e avaliava segundo os cinco critérios do ENEM, devidamente passados no prompt enviado pela API.

A plataforma retornou as notas por cada critério e a aplicação processava os resultados, gerando outro arquivo JSON atualizado com as notas, sendo depois gerado um arquivo do tipo CSV com os dados, para permitir sua análise comparativa. Esta etapa foi repetida para cada uma das quatro plataformas analisadas.

5. Análise de resultados. Os resultados apresentados por cada plataforma foram consolidados em uma única tabela de dados e analisados, usando a ferramenta estatística R para os testes e análises comparativas, por meio de estatísticas descritivas e inferenciais.

A tabela final com os resultados de todas as redações, nas quatro plataformas e por cada um dos cinco critérios do ENEM está disponibilizada no seguinte link: https://bit.ly/baseredacoesnota1000.

3. Resultados e Análises

Nessa seção, são apresentados e discutidos os resultados da pesquisa, conduzida conforme o percurso e materiais descritos anteriormente. Por causa da limitação do tamanho do texto, serão destacados somente os principais resultados obtidos.

3.1. Análise Descritiva

Considerando a natureza probabilística dos modelos de linguagem de grande escala (LLM), é reconhecido que respostas geradas por plataformas de IA generativa podem apresentar variações quando submetidas a múltiplas interações, mesmo sob condições controladas de prompt e dados de entrada.

Com o objetivo de mitigar esse efeito e conferir maior robustez aos resultados apresentados, o experimento de correção das redações não foi realizado em uma única execução. Cada redação foi submetida a cinco (05) correções independentes em cada uma das quatro plataformas avaliadas, mantendo-se constantes os *prompts*, critérios de avaliação e parâmetros de entrada. Para fins de análise estatística e apresentação dos resultados deste estudo, foi considerada a média das notas atribuídas em cada critério e da nota total, após as cinco interações por plataforma.

Esse procedimento permitiu reduzir oscilações pontuais decorrentes da variabilidade algorítmica, preservando tendências consistentes de desempenho entre as plataformas. Dessa forma, embora os resultados sejam fixos no contexto deste estudo, eles representam valores médios estáveis, o que reforça a confiabilidade, a reprodutibilidade e a relevância científica da pesquisa, mesmo diante de sistemas cuja saída não é determinística.

Os dados das notas de cada redação nas quatro plataformas foram disponibilizados em arquivo único contendo o ano, a nota total e as notas de cada critério, em cada uma das plataformas analisadas. A Tabela 1 apresenta as estatísticas descritivas básicas de cada um dos ambientes de LLM usados na pesquisa.

Tabela 1 – Estatísticas descritivas das notas de cada plataforma

Nota Geral	DeepSeek-r1	Gemini 1.5 Pro	GPT-4o-mini	Maritaca Sabiá3
Média	870,99	879,85	843,07	926,63
Mediana	860,00	880,00	850,00	930,00
Nota Máxima	960,00	980,00	880,00	960,00
Nota Mínima	720,00	660,00	720,00	820,00
Desvio-padrão	56,54	58,59	32,23	24,38
Erro Mínimo	4,00%	2,00%	12,00%	4,00%
Erro Máximo	28,00%	34,00%	28,00%	18,00%

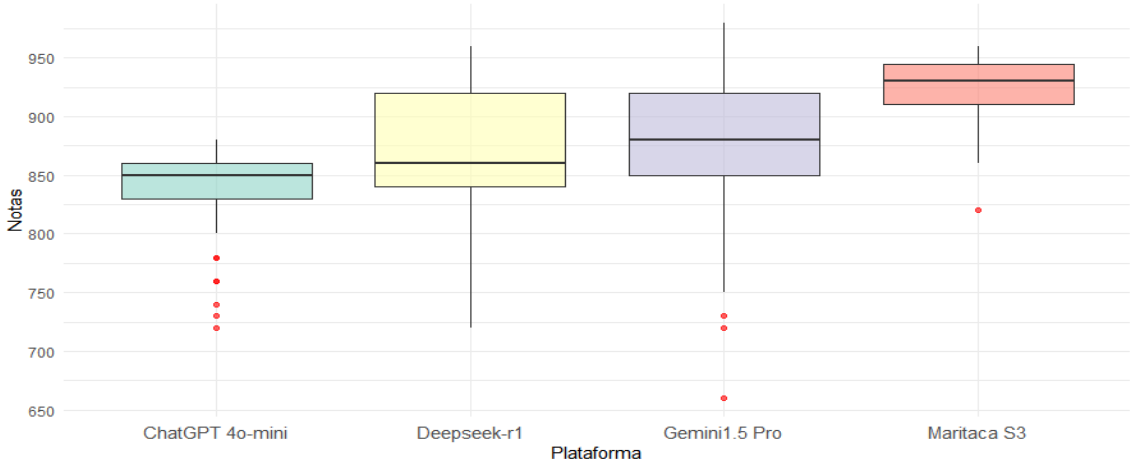
Fonte: Dados da pesquisa (2025).

Os resultados apontam um desempenho superior do Maritaca, na maioria dos indicadores apresentados, com uma menor diferença em relação à correção humana, com destaque a média, mediana e o menor erro máximo cometido. O erro máximo refere-se à maior diferença entre a nota 1000 e a menor nota dada pela plataforma, enquanto que o erro mínimo aponta para a diferença entre a nota 1000 e a maior nota dada pela plataforma.

Outro destaque refere-se ao Gemini, que apresentou nota máxima 980, com erro mínimo de 2,0%, o melhor desempenho nesses dois indicadores. Observa-se também que nenhuma redação obteve a nota 1000 em qualquer das plataformas. Isso pode sugerir algo relacionado com as bases textuais utilizadas para treinar os modelos das plataformas.

O Gráfico 1 apresenta visualmente a distribuição das notas de cada plataforma, em destaque as respectivas medianas e a presença de alguns outliers nos dados em três das plataformas. Observa-se também a menor variabilidade dos dados para o ChatGPT e o Maritaca, conforme também verificado nos valores seus desvios-padrão.

Gráfico 1 – *Boxplot* com distribuição das notas por plataforma**.**



Fonte: elaborado pelos autores

Ao serem analisadas as notas dadas em cada um dos 5 critérios da correção das redações, a plataforma Maritaca também obteve um desempenho superior em relação às demais, porém com alguns outros achados importantes nesta pesquisa. A Tabela 2 apresenta os resultados para os **Critérios 1 (Domínio da norma padrão da língua portuguesa) e 2 (Compreensão da proposta de redação)**.

Tabela 2 – Estatísticas das notas de cada plataforma, para os Critérios 1 e 2.

Critério 1	Deep	Gemin	ChGPT	Marit	Critério 2	Deep	Gemin	ChGPT	Marit
Média	177,52	167,18	176,14	181,68	Média	186,34	185,35	193,17	195,05
Nota Máx	200,00	190,00	180,00	190,00	Nota Máx	200,00	200,00	200,00	200,00
Nota Mín	160,00	100,00	150,00	160,00	Nota Mín	160,00	140,00	160,00	180,00
Desv Pad	8,72	15,49	8,08	5,28	Desv Pad	12,49	11,13	11,68	6,07
Erro Mín	0,00%	5,00%	10,00%	5,00%	Erro Mín	0,00%	0,00%	0,00%	0,00%
Erro Máx	20,00%	50,00%	25,00%	20,00%	Erro Máx	20,00%	30,00%	20,00%	10,00%

Fonte: Dados da pesquisa (2025).

Mesmo não apresentando nenhuma redação com nota máxima no critério 1, a Maritaca ainda assim obteve melhor média e desvio-padrão. Destaque também para a DeepSeek, como a única que atribuiu nota máxima nesse critério para as redações. Chama atenção o fato da Gemini ter tido uma única redação nota mínima 100, com 50% de erro máximo cometido, bastante relevante e sem uma razão aparente, já

que nas demais plataformas, a notas atribuídas para este critério na mesma redação foram 160 e 180 pontos.

Para o Critério 2, percebe-se um desempenho muito similar entre a Maritaca e o ChatGPT em relação à média das notas, entretanto o desvio-padrão da Maritaca foi quase a metade da sua concorrente, indicando uma variabilidade menor nas notas atribuídas. Outro ponto a ser destacado é que todas as plataformas deram a nota máxima nesse critério para algumas das redações, por consequência o erro mínimo foi zerado nesse critério nas plataformas.

A Tabela 3 exibe os resultados para os **critérios 3 (Seleção e organização de informações) e 4 (Demonstração de argumentação consistente)** da correção das redações.

Tabela 3 – Estatísticas das notas de cada plataforma, para os Critérios 3 e 4.

Critério 3	Deep	Gemin	ChGPT	Marit	Critério 4	Deep	Gemin	ChGPT	Marit
Média	166,34	170,94	163,27	181,44	Média	169,80	178,07	170,20	182,48
Nota Máx	200,00	200,00	180,00	195,00	Nota Máx	200,00	200,00	190,00	195,00
Nota Mín	140,00	120,00	140,00	160,00	Nota Mín	140,00	140,00	140,00	160,00
Desv Pad	13,91	14,21	7,33	8,57	Desv Pad	16,83	13,95	11,77	6,16
Erro Mín	0,00%	0,00%	10,00%	2,50%	Erro Mín	0,00%	0,00%	5,00%	2,50%
Erro Máx	30,00%	40,00%	30,00%	20,00%	Erro Máx	30,00%	30,00%	30,00%	20,00%

Fonte: Dados da pesquisa (2025).

Em relação ao Critério 3, observa-se o ChatGPT com menor variabilidade nas notas, seguida pela Maritaca. Entretanto, ambas não atribuíram nota máxima a nenhuma redação, sendo que a Maritaca apresentou notas com uma pontuação com variação ou intervalos de 5 pontos, o que não foi observado nas demais plataformas. O Gemini mais uma vez apresentou um alto erro máximo, com diferença substancial em relação às demais.

Para o Critério 4, mais uma vez houve uma prevalência da Maritaca embora, mais uma vez, a plataforma não tenha atribuído nota 200 neste critério, a nenhuma das 100 redações analisadas.

Por fim, o Critério 5 (**Elaboração de proposta de intervenção respeitando os direitos humanos**), que apresentou particularidades na correção das notas, em particular pelo Gemini e pelo ChatGPT, conforme mostra a Tabela 4.

Tabela 4 – Estatísticas das notas de cada plataforma, para o Critério 5.

Critério 5	Deep	Gemin	ChGPT	Marit
Média	170,99	177,43	140,30	185,99
Nota Máx	200,00	200,00	180,00	195,00
Nota Mín	120,00	0,00	0,00	150,00
Desv Pad	19,63	23,30	25,27	7,74
Erro Mín	0,00%	0,00%	10,00%	2,50%
Erro Máx	40,00%	100,00%	100,00%	25,00%

Fonte: Dados da pesquisa (2025).

Observa-se que duas plataformas atribuíram nota 0 (zero), nesse critério, para redações. Analisando a base de dados, constatou-se que somente duas redações distintas foram zeradas, uma em cada plataforma. O critério relacionado à proposta de intervenção com respeito aos direitos humanos pode ter levado a essa nota. Ao solicitar nova avaliação das redações e pedir a inclusão de feedbacks com a justificativa da nota dada, obteve-se o seguinte retorno:

Gemini – Redação com o tema *"**Manipulação do comportamento do usuário pelo controle de dados na internet**"*, ano 2018:

Segundo o feedback dado pela plataforma, "*a redação analisada, embora bem estruturada e argumentativa, não apresenta uma proposta de intervenção clara e detalhada para o problema da manipulação do comportamento do usuário pelo controle de dados na internet. A redação discute o problema de forma eficaz, mas não sugeriu soluções concretas, como leis, ferramentas de controle de dados ou campanhas de conscientização, que são necessárias para uma boa pontuação neste critério*."

ChatGPT – Redação com o tema "***Publicidade infantil em questão no Brasil***", ano 2014:

"A nota zero no critério 5 ocorreu porque a proposta de intervenção não apresentou todos os elementos exigidos pelo ENEM: agente, ação, modo/meio, efeito e detalhamento. Embora a autora tenha sugerido que o governo e as escolas sejam responsáveis por medidas como a criação de leis rígidas e oficinas educativas, a explicação de como essas ações seriam implementadas não foi suficientemente detalhada.

Além disso, a proposta foi considerada vaga e pouco viável, o que compromete sua efetividade na resolução do problema. Como consequência, a redação não atendeu plenamente ao critério de elaboração de uma intervenção concreta e estruturada, resultando na nota zero."

Como foram somente dois casos em cem redações, cada uma com cinco critérios, observou-se a ocorrência de 0,4% de notas zeros que foram atribuídas pelas plataformas, indicador que pode ser considerado baixo em razão do número total de redações corrigidas (100 redações x 5 critérios = 500 correções).

De uma forma geral, a Maritaca mais uma vez obteve bons indicadores estatísticos. Entretanto, nesse critério, a plataforma apresentou o seu maior erro máximo entre todos os critérios de correção, o que sugere um nível de complexidade maior desse critério em relação aos demais.

Finalizando as análises descritivas, foi feita também uma análise das médias das notas das redações agrupadas por ano de aplicação, para proporcionar uma avaliação mais horizontal das correções, por se tratar de comparação de redações sob o mesmo tema. Os resultados dessa comparação estão na Tabela 5. Em destaque, a maior média das notas totais atribuída pela plataforma em todo o período analisado.

Tabela 5 – Média das notas de cada plataforma, por cada ano.

Ano	DeepSeek-r1	Gemini 1.5 Pro	GPT-4o-min	Maritaca S3
2014	832,00	819,00	820,00	907,50
2015	833,00	872,00	839,00	919,00
2016	883,08	***905,77***	833,85	930,77
2017	851,25	858,75	856,25	921,25
2018	889,29	876,43	844,29	932,14
2019	892,00	889,00	***858,00***	930,50
2020	884,29	904,29	845,71	926,43
2021	866,25	893,75	837,50	***936,25***
2022	***908,00***	890,00	***858,00***	933,00
2023	860,00	888,18	842,73	926,82

Fonte: Dados da pesquisa (2025).

Os resultados por ano apontam que a média das notas atribuídas pela plataforma Maritaca sempre foram acima de 900, indicando valores mais próximos à correção dos especialistas em todos os anos. Ressalta-se mais uma vez que, mesmo com melhor desempenho, a Maritaca apresentou nota máxima (200) somente no critério 2, para algumas das redações. Em outro sentido, o ChatGPT não apresentou nenhuma média acima de 900 pontos e tem como valores máximos 858 pontos, média das notas atribuídas para redações de 2019 e 2022. Nas outras duas plataformas, a ocorrência de médias acima de 900 também é muito baixa.

Percebe-se também que essas maiores médias foram distribuídas ao longo dos anos, não tendo nenhum ano cujas médias nas quatro plataformas tenham sido maiores em relação aos demais anos, o

que pode caracterizar uma uniformidade menor nos modelos usados para treinamento nas plataformas, além de diferentes parâmetros e ajustes dos modelos.

Um aspecto relevante para a interpretação dos resultados observados diz respeito à desigualdade linguística presente nos dados de treinamento dos modelos de linguagem de grande escala. Estudos recentes têm demonstrado que a maior parte dos *corpora* utilizados no treinamento desses modelos é composta predominantemente por textos em língua inglesa e provenientes de contextos socioculturais do Norte Global, o que pode limitar o desempenho dos sistemas em tarefas que demandam compreensão aprofundada de outros idiomas e de contextos locais específicos (Longpre et al., 2024).

À luz desse cenário, o desempenho superior da plataforma Maritaca IA pode estar associado à hipótese de que seus modelos foram treinados com maior predominância de textos em língua portuguesa e com dados mais alinhados ao contexto educacional e sociocultural brasileiro.

Essa característica pode ter favorecido uma melhor interpretação dos critérios oficiais de correção do ENEM, bem como maior sensibilidade às nuances linguísticas, argumentativas e discursivas presentes nas redações analisadas, contribuindo para a menor distância observada em relação às avaliações realizadas por especialistas humanos.

3.2. Testes estatísticos para verificação de diferenças significativas

Apesar de as análises das estatísticas descritivas apresentarem indicativos importantes sobre o desempenho das plataformas na correção das redações que obtiveram nota 1000 no ENEM, buscou-se confirmar os resultados a partir da aplicação dos testes estatísticos ANOVA e Tukey, para verificar se existem diferenças significativas nas avaliações das quatro plataformas analisadas e, caso positivo, em quais e em que valor seriam essas diferenças.

Antes da aplicação do teste de Análise de Variância (ANOVA), foram avaliados os principais pressupostos necessários à sua utilização, conforme recomenda a literatura estatística. A normalidade da distribuição dos resíduos foi verificada por meio do teste de Shapiro-Wilk, complementado por inspeção visual de histogramas, não sendo identificadas violações graves que comprometessem a aplicação do teste paramétrico.

O teste ANOVA (Análise de Variância) é um método estatístico utilizado para comparar as médias de três ou mais grupos e verificar se há diferenças estatisticamente significativas entre eles. Ele avalia a variabilidade entre os grupos, baseando-se na razão F, que mede a relação entre a variância explicada (entre grupos) e a variância não explicada (dentro dos grupos). Se o valor-p resultante for menor que um nível de significância pré-definido (neste caso, 0,05), rejeita-se a hipótese nula de que todas as médias são iguais (Montgomery, 2017). Embora o ANOVA informe se há uma diferença entre os grupos, ele não identifica quais grupos diferem entre si. Para isso, testes post hoc, como o teste de Tukey, são frequentemente utilizados.

O teste de Tukey, também conhecido como *Tukey's Honest Significant Difference* (HSD), é um teste utilizado após uma ANOVA para identificar quais pares de médias diferem significativamente entre si. O teste calcula um intervalo de confiança para cada diferença entre pares de médias e considera diferenças estatisticamente significativas se esses intervalos não contiverem zero (Abdi & Williams, 2010).

Para realização dos testes, foram criadas novas colunas de dados com as respectivas diferenças de cada nota total (**Diferença da Nota Total = 1000 – nota total dada pela plataforma**) e para cada critério (**Diferença da Nota no Critério = 200 – nota no critério dada pela plataforma**), que foram atribuídas em cada plataforma. Assim, os testes foram feitos usando os dados das diferenças de notas total e por critérios em cada redação, nas quatro plataformas. Como foram usadas as diferenças entre notas total e por critério, é importante observar que:

- **Valores menores nas diferenças indicam notas mais altas atribuídas pela plataforma (melhor desempenho).**

- **Valores maiores nas diferenças indicam notas mais baixas atribuídas pela plataforma (pior desempenho).**

O teste ANOVA avaliou se houve diferenças estatisticamente significativas entre as médias das diferenças para as quatro plataformas: Gemini, DeepSeek-r1, GPT-4o-mini e Maritaca. O teste apontou um valor de F= 58.4 e um p-valor: < 2e-16 (Altamente significativo). Isso indica que pelo menos uma das plataformas atribuiu notas significativamente diferentes das demais.

O Teste de Tukey comparou as diferenças entre as notas concedidas por cada par de plataformas. Só houve uma ocorrência sem diferença estatisticamente significativa (p > 0.05), indicando que ambas tiveram desempenhos similares, atribuindo notas próximas, com as notas ligeiramente mais altas para o Gemini (Gemini vs. DeepSeek-r1, p-valor = 0.51, diferença = -8.86). As demais comparações são apresentadas no Quadro 1, todas com diferenças significativas (P =< 0.0001).

Quadro 1 – Resultados do Teste de Tukey entre as plataformas.

Plataf. 1	Plataf. 2	Diferença	Conclusão
Maritaca	DeepSeek	-55.64	O alto valor negativo indica que o Maritaca atribuiu notas significativamente mais altas do que o DeepSeek
Maritaca	Gemini	-46.78	Maritaca atribuiu notas significativamente mais altas do que o Gemini.
Maritaca	ChatGPT	-83.56	Maritaca atribuiu notas significativamente mais altas do que o ChatGPT.
ChatGPT	DeepSeek	27.92	O valor positivo indica que o ChatGPT atribuiu notas significativamente mais baixas do que o DeepSeek.
ChatGPT	Gemini	36.78	ChatGPT atribuiu notas significativamente mais baixas do que o Gemini.

Fonte: Dados da pesquisa (2025).

Esses resultados dos testes estatísticos corroboraram com as análises descritivas apresentadas na seção anterior, com a Maritaca obtendo o melhor desempenho e o ChatGPT em pior colocação entre as quatro plataformas analisadas.

Ao serem buscadas explicações e justificativas para o melhor desempenho da plataforma Maritaca nas correções das redações, aproximando-se mais das notas dadas pelos especialistas do que as demais plataformas, a suposição mais relevante foi a de que o fato da plataforma utilizar, para treinamento dos seus modelos, dados públicos em sua maioria em Português. Como a plataforma foi treinada especificamente para entender bem a língua portuguesa, portanto, espera-se que tenha um desempenho superior em tarefas nesse idioma (Maritaca, 2025).

Um recente estudo da *Data Provenance Iniciative* (Longpre, 2024) apresentou uma auditoria sobre a proveniência de dados usados no treinamento de modelos de IA multimodal, abrangendo quase 4.000 conjuntos de dados públicos de 1990 a 2024. O estudo analisou três modalidades principais de fontes de dados: texto, fala e vídeo, destacando tendências de obtenção de dados, restrições de uso e representatividade geográfica e linguística.

Apesar de o aumento absoluto no número de línguas representadas (608) e países (67), a distribuição continua concentrada em regiões ocidentais. Segundo o estudo, 93% dos *tokens* de texto vêm de organizações da América do Norte e Europa, enquanto África e América do Sul representam menos de 0,2% dos dados. Essa escassez de dados para várias línguas prejudica o desenvolvimento de modelos de IA que atendam de forma mais equitativa às populações globais.

Esse estudo reforça a suposição que o fator idioma dos dados dos modelos pode ter influenciado no melhor desempenho da Maritaca frente às demais plataforma sem deixar de considerar outros aspectos técnicos importantes da plataforma.

4. Conclusões e Trabalhos Futuros

Este estudo trouxe uma análise comparativa relevante de quatro plataformas de IA Generativa, que utilizam LLM, na comparação de 100 redações que obtiveram nota máxima no ENEM, ao longo do período de 2014 a 2023.

A pesquisa apontou que a plataforma brasileira Maritaca foi a que mais se aproximou da correção feita por especialistas, obtendo os melhores resultados nas estatísticas descritivas básicas, confirmados pelo Teste ANOVA e Tukey. Em outro sentido, o ChatGPT, uma das plataformas pioneiras para uso geral da IA Generativa, foi a que teve o pior desempenho entre as quatro plataformas analisadas. O estudo trouxe ainda o Gemini, do Google e a DeepSeek.

A não atribuição da nota 1000 em nenhuma redação foi uma observação importante, assim como a nota 0 no critério 5 para duas redações, dadas por duas das plataformas. Isso pode ser resultado ainda de ajustes nos parâmetros dos modelos ou mesmo das bases usadas nas plataformas, na qual ainda predomina textos em outros idiomas, exceto na Maritaca. Destaca-se ainda que foram usados somente as versões gratuitas das plataformas, o que pode também ter afetado o processo de correção, com bases menos atualizadas e modelos com menos parâmetros.

Há também o fato que podem ocorrer variações nas notas cada vez que é feita uma nova solicitação de correção pelas plataformas, mas estas pequenas mudanças não inviabilizam o presente estudo, já que não alteraram significativamente as notas, nem provocaram mudanças na ordem de desempenho das plataformas.

A análise apresentada neste estudo reforça a necessidade de uma abordagem crítica e responsável quanto ao uso da IA no campo educacional, especialmente em atividades relacionadas à produção e à avaliação textual. Embora as plataformas de IA generativa demonstrem potencial significativo para analisar padrões linguísticos, organizar critérios avaliativos e oferecer *feedback* estruturado, seu uso levanta dilemas importantes de natureza ética, pedagógica e formativa, como os limites da automatização, a transparência dos processos algorítmicos, a dependência tecnológica e a preservação da autoria humana, tanto na escrita quanto na revisão dos textos.

À luz dessa análise, a IA não se apresenta como substituta do trabalho humano, mas como um instrumento de apoio que pode ampliar capacidades, otimizar processos e auxiliar estudantes e educadores no desenvolvimento de competências linguísticas e avaliativas, desde que utilizada de forma crítica, mediada e consciente. Assim, o valor educacional da IA reside menos na automação da correção ou da autoria e mais em sua contribuição como ferramenta auxiliar para a aprendizagem, a reflexão e o aprimoramento das práticas humanas de escrita e avaliação.

Como trabalhos futuros, este estudo está atrelado a um projeto de iniciação tecnológica que pretende desenvolver uma ferramenta, em parceria com uma escola pública, para apoiar os estudantes na escrita e os professores na correção das redações, agora com a possibilidade de adoção da plataforma Maritaca como base para o projeto.

Por fim, ressalta-se que a capacidade da escrita é um importante componente na formação cultural e profissional das pessoas e sempre vai existir a necessidade de um refinamento humano nos textos gerados por IA, sejam eles de qualquer natureza ou temática.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (**CNPq**) e à Univasf pela concessão de Bolsa de Iniciação Tecnológica e ao Colégio Rui Barbosa pelo apoio para realização deste estudo.

Referências Bibliográficas

ABDI, H.; WILLIAMS, L. J. Tukey's Honestly Significant Difference (HSD) Test. In: Salkind, N. J. (Ed.), **Encyclopedia of Research Design**. SAGE Publications. 2010.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). **Cartilha do Participante: Redação no ENEM**. Brasília, 2024. Disponível em https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/a_redacao_no_enem_2024_cartilha_do_participante.pdf . Acesso em 02 fev. 2025.

CASTELLI, M.; MANZONI, L. Generative models in artificial intelligence and their applications. **Applied Sciences**, v. 12, n. 9, p. 4127, 2022.

CHATGPT. **GPT-4 Technical Report**. https://arxiv.org/html/2303.08774v6 . Acesso em 02 fev. 2025.

COSTA, M. P. F. *et al*. Panorama histórico das propostas de redações do ENEM: um olhar sobre as temáticas e critérios de avaliação. **Revista Diálogo Educacional**, v. 23, n. 78, p. 1332-1352, 2023.

DEEPSEEK. **DeepSeek-v3 technical report**. 2024. Disponível em https://arxiv.org/html/2412.19437v1 . Acesso em 02 fev. 2025.

EPSTEIN, Z. *et al.* Art and the science of generative AI. **Science**, v. 380, n. 6650, p. 1110-1111, 2023. Disponível em https://arxiv.org/abs/2306.04141 . Acesso em 10 fev. 2025.

GEMINI. **Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context**. 2024. Disponível em https://arxiv.org/abs/2403.05530 . Acesso em 02 fev. 2025.

KASNECI, E. et. al. ChatGPT for good? On opportunities and challenges of large language models for education. **Learning and Individual Differences**, v. 103, [s. n.], [s. p.], mar. 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S1041608023000195 . Acesso em 28 mar. 2025.

KEHOE, F. Leveraging Generative AI Tools for Enhanced Lesson Planning in Initial Teacher Education at Post Primary. **Irish Journal of Technology Enhanced Learning**, 7(2), 172-182. 2023. Disponível em: https://doi.org/10.22554/ijtel.v7i2.124 . Acesso em 25 fev. 2025.

LONGPRE, S. *et al.* **Bridging the Data Provenance Gap Across Text, Speech and Video**. 2025. Disponível em: https://arxiv.org/abs/2412.17847. Acesso em 23 mar. 2025.

MARITACA. **Maritaca AI**. Disponível em https://www.maritaca.ai/sobre-maritaca-ai. Acesso em 18 fev. 2025.

MONTGOMERY, D. C. **Design and analysis of experiments**. John wiley & sons, 2017.

PERES, R. *et al.* On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. **International Journal of Research in Marketing**, v. 40, n. 2, p. 269-275, 2023.

SABIÁ. **Sabiá 3 Technical Report**. Disponível em https://arxiv.org/pdf/2410.12049 . Acesso em 18 fev. 2025.

SALLAM, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. **Healthcare**, v. 11, n. 6, p. 887, 2023. Disponível https://www.mdpi.com/2227-9032/11/6/887 . Acesso em 25 mar. 2025.