INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
CAMPUS CAMPINA GRANDE
BACHARELADO EM ENGENHARIA DA COMPUTAÇÃO

ELIEL ALEXANDRE ANDRADE SOUZA

AVALIAÇÃO INTELIGENTE DE REDAÇÕES DO ENEM: UM ESTUDO COMPARATIVO DOS GRANDES MODELOS DE LINGUAGEM

CAMPINA GRANDE, PARAÍBA
2025

ELIEL ALEXANDRE ANDRADE SOUZA

AVALIAÇÃO INTELIGENTE DE REDAÇÕES DO ENEM: UM ESTUDO COMPARATIVO DOS GRANDES MODELOS DE LINGUAGEM

Trabalho de Conclusão de Curso (monografia) apresentado como exigência parcial para obtenção do diploma do Bacharel em Engenharia da Computação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, Campus Campina Grande.

Orientador: Prof. DSc. Danyllo Wagner Albuquerque

CAMPINA GRANDE, PARAÍBA
2025

Catalogação na fonte:
Ficha catalográfica elaborada por Gustavo César Nogueira da Costa - CRB 15/479

S729a Souza, Eliel Alexandre Andrade.

Avaliação inteligente de redações do ENEM: um estudo comparativo dos grandes modelos de linguagem / Eliel Alexandre Andrade Souza. - Campina Grande, 2025.
27 f.: il.

Trabalho de Conclusão de Curso (Graduação em Bacharelado em Engenharia de Computação) - Instituto Federal da Paraíba, 2025.
Orientador: Prof. Danyllo Wagner Albuquerque, DSc.

1. Modelos de linguagem. 2. Inteligência artificial. 3. Avaliação de textos. 4. ENEM – Exame Nacional do Ensino Médio. I. Albuquerque, Danyllo Wagner. II. Título.

CDU 004.8

ELIEL ALEXANDRE ANDRADE SOUZA

AVALIAÇÃO INTELIGENTE DE REDAÇÕES DO ENEM: UM ESTUDO COMPARATIVO DOS GRANDES MODELOS DE LINGUAGEM

Trabalho de Conclusão de Curso (monografia) apresentado como exigência parcial para obtenção do diploma do Bacharel em Engenharia da Computação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, Campus Campina Grande.

Aprovada em _____/_____/_____.

BANCA EXAMINADORA:

Prof. DSc. Danyllo Wagner Albuquerque(Orientador)
Instituto Federal da Paraíba (IFPB)

Prof. DSc. Marcelo José Siqueira Coutinho de Almeida
Instituto Federal da Paraíba (IFPB)

Prof. DSc. Igor Barbosa da Costa
Instituto Federal da Paraíba (IFPB)

CAMPINA GRANDE, PARAÍBA
2025

Dedico este trabalho a todos que acreditaram que ele sairia.

AGRADECIMENTOS

Agradeço primeiramente a Deus pela sua infinita graça e misericórdia. Aos meus pais por todo o amor, cuidado e carinho, e por serem a base do que eu sou hoje. Aos meus mais íntimos colegas de curso, Jonas e Jorge, que tornaram a caminhada do curso mais leve. À Ana Clara, por ser a mulher mais incrível que eu conheci na vida. Por fim, a todos os professores e servidores que sustentam a instituição e são peças fundamentais na formação dos profissionais.

"Não espere o futuro mudar sua vida
Porque o futuro será a consequência do presente".
Racionais MC's

RESUMO

A correção de redações do ENEM é um processo que exige análise de competências textuais, tornando-se um desafio em termos de tempo e consistência avaliativa. Com o avanço da Inteligência Artificial e o surgimento de Grandes Modelos de Linguagem (GMLs), têm-se a oportunidade de automatizar esse processo de modo eficiente. Este estudo investiga o uso de GMLs para correção automática de redações do ENEM, analisando a diferença de desempenho entre modelos treinados e não-treinados. Para isso, foi conduzido um experimento comparando o desempenho do GPT com um corretor humano, analisando 40 redações do ENEM distintas. Os resultados indicam que o GPT treinado alcançou uma taxa de acerto de até 50%, enquanto a versão não-treinada obteve, em média, 30% de precisão. Isso sugere que o GPT, após treinamento, conseguiu maior alinhamento com os critérios do ENEM. Conclui-se que a aplicação de GMLs na correção de redações do ENEM tem potencial para aumentar a eficiência e padronização das avaliações. Contudo, ajustes no treinamento e calibração das notas ainda são necessários para melhorar a confiabilidade do sistema.

Palavras-chaves: GML; Inteligência Artificial; ENEM; GPT.

ABSTRACT

The correction of ENEM essays is a process that requires a thorough analysis of textual competencies, making it a challenge in terms of time and evaluation consistency. With the advancement of Artificial Intelligence and the emergence of Large Language Models (LLMs), there is an opportunity to automate this process efficiently. This study investigates using LLMs to automatically grade ENEM essays, analyzing the performance differences between trained and untrained models. A comparative experiment was conducted comparing GPT and human, evaluating 40 distinct essays. The results indicate the trained GPT achieved an accuracy rate of up to 50%, while the untrained version reached an average accuracy of 30%. This suggests that GPT, after training, aligned more closely with ENEM criteria. The study concludes that LLMs have the potential to enhance efficiency and standardization in essay grading, but further training adjustments and score calibration are necessary to improve system reliability.

Keywords: LLM; Artificial Intelligence; ENEM; GPT.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

GML	Grande Modelo de Linguagem
ENEM	Exame Nacional do Ensino Médio
GPT	Generative Pre-trained Transformer
IFPB	Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
TIC	Tecnologias da Informação e Comunicação
PLN	Processamento de Linguagem Natural
OCR	Reconhecimento Óptico de Caracteres

SUMÁRIO

1 INTRODUÇÃO

Os Grandes Modelos de Linguagem (GML) representam um avanço significativo na forma como máquinas processam e interpretam a linguagem humana, utilizando redes neurais profundas treinadas com vastos conjuntos de dados (LEE, 2023). Esses modelos têm demonstrado um desempenho notável em diversas aplicações, como geração de texto, resumo automático e tradução, transformando significativamente áreas como comunicação, pesquisa e educação (GUIMARÃES *et al.*, 2023).

No contexto educacional, os GMLs têm sido explorados para otimizar processos de ensino e aprendizagem, oferecendo suporte a alunos e professores na produção de textos, revisão gramatical e aprimoramento da escrita (LIMA, 2023) (MONTEIRO, 2023). No entanto, sua aplicação na avaliação de redações ainda enfrenta desafios, especialmente quanto à fidelidade das correções aos critérios avaliativos e à consistência na atribuição de notas. A padronização da correção de redações do Exame Nacional do Ensino Médio (ENEM), por exemplo, exige rigor na análise de competências textuais, tornando a automação desse processo um desafio complexo.

Diante disso, este estudo investiga a viabilidade do uso de GMLs na correção automática de redações do ENEM, comparando o desempenho de modelos treinados e não-treinados na atribuição de notas e *feedbacks*. Para isso, foi conduzido um experimento com 40 redações, avaliadas pelo modelo GPT-4 treinado e não-treinado, analisando sua capacidade de fornecer avaliações coerentes com os critérios oficiais do exame.

O presente estudo contribui para o entendimento do papel dos GMLs na avaliação textual automatizada. Primeiramente, ele analisa a eficácia dos modelos em diferentes estágios de treinamento, destacando seus impactos na qualidade da correção. Além disso, discute os desafios enfrentados na adaptação desses modelos aos critérios do Exame Nacional do Ensino Médio (ENEM), considerando questões como coerência na pontuação, limitações na interpretação textual e possíveis vieses algorítmicos. Por fim, a pesquisa apresenta recomendações para aprimorar o uso dessas ferramentas, possibilitando sua integração mais eficaz no contexto educacional e contribuindo para um processo avaliativo mais eficiente e padronizado.

Finalmente, o artigo está estruturado da seguinte forma: após esta introdução, a Seção 2 discute a fundamentação teórica dos GMLs. A Seção 3 examina trabalhos relacionados para contextualizar nossa pesquisa dentro do campo existente. A Seção 4 detalha nossa metodologia de levantamento de dados. Nas Seções 5 e 6, apresentamos os resultados obtidos e discutimos suas limitações e ameaças, respectivamente. A Seção 7 aborda as considerações finais e sugestões de direções para pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

A aplicação de GMLs na educação tem despertado crescente interesse devido ao seu potencial de otimizar a produção e a correção textual. Essas tecnologias, baseadas em redes neurais profundas e aprendizado de máquina, vêm sendo empregadas em diversos contextos educacionais para melhorar a interação entre alunos e professores, bem como para aprimorar a análise e a avaliação de textos escritos (LEE, 2023).

No Brasil, um dos desafios na avaliação de textos está na correção da redação do ENEM, que adota um sistema de pontuação baseado em cinco competências. Cada competência recebe uma nota de 0 a 200 pontos, totalizando até 1000 pontos. Os critérios de avaliação incluem (1) domínio da norma culta da língua portuguesa, (2) compreensão e desenvolvimento do tema proposto, (3) capacidade de organização e argumentação coerente do texto, (4) domínio dos mecanismos linguísticos necessários à construção da coesão textual e (5) elaboração de uma proposta de intervenção para o problema abordado (INEP, 2023). A correção dessas redações é feita manualmente por dois avaliadores, podendo haver uma terceira correção caso haja discrepâncias significativas entre as notas atribuídas.

A evolução das Tecnologias da Informação e Comunicação (TIC) proporcionou avanços na gestão do conhecimento educacional, permitindo a automação de processos e a personalização da aprendizagem (GUZMAN *et al.*, 2022). No entanto, a avaliação de redações continua sendo um processo desafiador, devido à subjetividade inerente à análise textual. Diante disso, os GMLs emergem como uma alternativa promissora para auxiliar na correção automatizada de redações do ENEM, reduzindo inconsistências e proporcionando *feedbacks* detalhados aos candidatos.

Com base nessa perspectiva, modelos como GPT e Gemini têm sido explorados para atribuir notas e oferecer *feedback* automatizado, com o objetivo de padronizar e agilizar a avaliação de redações. Estudos recentes demonstram que esses modelos podem desempenhar um papel relevante na identificação de erros gramaticais (LIMA, 2023), na estruturação do texto e na coerência argumentativa (MONTEIRO, 2023), aspectos essenciais na avaliação do ENEM.

Entretanto, apesar dos avanços, a aplicação dos GMLs na correção de redações enfrenta desafios importantes. Primeiramente, a capacidade dos modelos de compreender a complexidade dos critérios avaliativos ainda precisa ser refinada, pois variações no treinamento podem impactar significativamente os resultados (VARAS *et al.*, 2023). Além disso, questões como viés algorítmico, dificuldades na interpretação de contextos subjetivos e a necessidade de calibração das notas exigem uma abordagem cuidadosa na implementação dessas ferramentas (CARNEIRO *et al.*, 2018).

Este estudo busca contribuir para esse debate, analisando a viabilidade do uso de GMLs na correção automática de redações do ENEM. Ao comparar modelos treinados e não-treinados, investigamos como essas ferramentas podem melhorar a coerência e a precisão das avaliações, ao mesmo tempo em que exploramos seus desafios e limitações no contexto educacional.

3 TRABALHOS RELACIONADOS

A correção automática de redações com modelos de Inteligência Artificial tem sido objeto de crescente investigação, principalmente com a adoção de GML. Diversos estudos exploram o potencial dessas ferramentas na avaliação textual, analisando aspectos como coerência, precisão gramatical e confiabilidade das pontuações geradas.

Um primeiro grupo de estudos investigou o desempenho do ChatGPT na produção e avaliação de textos. Fitria (2023) examina como o modelo gera redações em inglês, apontando que, embora produza textos coerentes, a precisão gramatical ainda precisa de melhorias, sugerindo a necessidade de refinamentos na modelagem linguística (FITRIA, 2023). Bui et al. (2024) avaliam a confiabilidade do ChatGPT na atribuição de notas, comparando suas avaliações com as de corretores humanos. O estudo revela que as pontuações geradas não se alinham bem às avaliações experientes e apresentam inconsistência entre múltiplas análises, evidenciando desafios na padronização da correção (BUI; BARROT, 2024). Por outro lado, Altamimi et al. (2023) analisam a eficácia do ChatGPT na automação da correção em diversas áreas acadêmicas, concluindo que o GPT-4 supera o GPT-3 em precisão e eficiência, demonstrando um potencial significativo para aprimorar métodos tradicionais de avaliação (ALTAMIMI, 2023).

Outro conjunto de estudos foca especificamente na avaliação automatizada de redações em português, com ênfase na aplicação de aprendizado de máquina e Processamento de Linguagem Natural (PLN). Lima et al. (2023) realizam uma revisão sistemática sobre os métodos utilizados na correção automática de textos em língua portuguesa, destacando as principais técnicas aplicadas e as dimensões mais frequentemente analisadas (LIMA *et al.*, 2023). Neto et al. (2020) investigam o uso de aprendizado de máquina para melhorar a qualidade das avaliações automatizadas em comparação com corretores humanos, apresentando resultados promissores (NETO *et al.*, 2020). Já Júnior et al. (2017) propõem um sistema de correção automática para redações do ENEM, avaliando especificamente a Competência 1 (domínio da norma culta da língua portuguesa) com resultados que imitam a avaliação humana com baixo erro médio absoluto (JÚNIOR; SPALENZA; OLIVEIRA, 2017).

Este estudo se diferencia ao analisar todas as cinco competências do ENEM, proporcionando uma avaliação mais completa do texto. Além disso, enquanto grande parte dos estudos foca em modelos estáticos, aqui são comparados modelos treinados e não-treinados, evidenciando o impacto do treinamento na precisão das avaliações. A análise também verifica a consistência das notas atribuídas pelos modelos GPT e Gemini, avaliando como a automação pode reduzir variações e melhorar a padronização da correção. Por fim, este estudo representa uma iniciativa pioneira na validação de

GMLs para a correção de redações do ENEM, contribuindo para o avanço da automação na avaliação textual em português e fornecendo subsídios para o desenvolvimento de novas ferramentas educacionais.

4 METODOLOGIA

Este estudo investiga a viabilidade de utilizar GMLs na correção automática de redações do ENEM, comparando o desempenho de modelos treinados e não-treinados na atribuição de notas e *feedback* textual. Para isso, foi conduzido um experimento controlado estruturado em quatro etapas principais: seleção do corpus (Seção 4.1), processamento dos dados (Seção 4.2), avaliação dos modelos (Seção 4.3) e análise dos resultados (Seção 4.4).

4.0.1 Seleção do *Corpus*

A primeira etapa consistiu na definição e coleta do conjunto de redações utilizado para a avaliação dos modelos. O *corpus* foi composto por 40 redações extraídas de fontes diversas, garantindo a inclusão de textos com diferentes níveis de proficiência na escrita (vide Material Suplementar (ALBUQUERQUE, 2025)). A seleção das redações considerou a diversidade temática e a variação na qualidade textual, com o objetivo de analisar o comportamento dos modelos em diferentes cenários avaliativos. A amostragem contemplou redações com características que vão desde textos bem estruturados e linguisticamente refinados até produções com falhas na coesão, coerência e domínio da norma culta.

Após a coleta, as redações foram inicialmente digitalizadas utilizando ferramentas de Reconhecimento Óptico de Caracteres (OCR), possibilitando a conversão de textos manuscritos em formato editável. Esse procedimento permitiu a utilização de redações em formato físico no experimento, garantindo uma maior diversidade no *corpus*. Em seguida, os textos foram avaliados individualmente para assegurar corretude e completude, garantindo que todas as informações relevantes fossem preservadas.

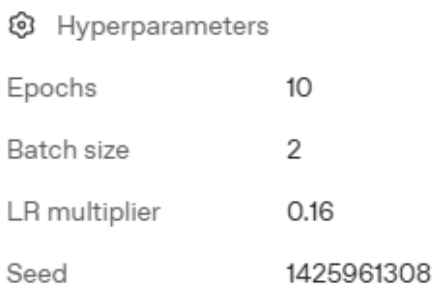
4.0.2 Processamento dos Dados

Após a digitalização e validação, as redações passaram por um processo de normalização textual para garantir uniformidade na entrada dos modelos. Foram removidos ruídos nos dados, como caracteres especiais, formatações inconsistentes e espaços em excesso, de modo a minimizar interferências na interpretação dos textos pelos modelos. A padronização foi realizada utilizando ferramentas de pré-processamento linguístico, assegurando que todas as redações estivessem em um formato homogêneo antes da aplicação dos modelos de correção.

Em seguida, os textos foram submetidos à avaliação por duas variantes dos modelos de linguagem: GPT não-treinado e GPT treinado. O processo de treinamento envolveu ajustes finos (*fine-tuning*) utilizando técnicas de aprendizado supervisionado,

em que os modelos foram expostos a exemplos rotulados de correções humanas para otimizar a atribuição de notas e a produção de *feedbacks* coerentes. Esses exemplos de correções ficaram compilados em um arquivo *.jsonl*, submetido ao processo de *fine-tuning* disponível na plataforma da própria OpenAI. O arquivo de treinamento pode ser visto em um Material Suplementar (ANDRADE, 2025), e os parâmetros utilizados no treinamento podem ser vistos na Figura 1. Adicionalmente, utilizou-se documentação oficial do INEP contendo informações para correção e pontuação de cada uma das cinco competências, permitindo que os modelos ajustassem sua parametrização para melhor aderência aos critérios de avaliação do ENEM.

Figura 1 – Parâmetros utilizados para *fine-tuning*.



A escolha do modelo GPT(na versão “gpt-4o”) para o experimento foi motivada por três fatores principais. O primeiro fator diz respeito à capacidade de processamento de linguagem natural (PLN) desses modelos, ambos reconhecidos pelo desempenho avançado na geração e análise de textos. O segundo fator está relacionado à disponibilidade e suporte técnico, uma vez que o GPT possui versões treináveis, ampla documentação e APIs acessíveis, permitindo ajustes mais refinados para a tarefa específica de correção de redações. O terceiro fator envolve estudos prévios e *benchmarks*, que indicam que esses modelos são mais robustos para avaliações complexas em comparação com alternativas como LLaMA ou Claude, que apresentam limitações em contextualização e coerência na correção textual.

4.0.3 Avaliação dos Modelos

A avaliação dos modelos foi conduzida com base nos cinco critérios estabelecidos pelo ENEM, garantindo uma análise detalhada das competências exigidas na prova. Cada redação foi corrigida de forma independente por cada versão dos modelos, sendo atribuídas notas que variam de 0 a 200 pontos para cada competência, totalizando um máximo de 1000 pontos. Os critérios de avaliação encontram-se descritos na Tabela 1.

Para garantir a reprodutibilidade e a confiabilidade dos resultados, as notas atribuídas pelos modelos foram comparadas entre si e confrontadas com as diretrizes oficiais de avaliação do exame. Além disso, foram gerados relatórios de *feedback* textual para cada redação, permitindo uma análise qualitativa da capacidade dos

Tabela 1 – Competências avaliadas na redação do ENEM

Competência	Descrição
1	Demonstrar domínio da norma culta da língua portuguesa na modalidade escrita formal.
2	Compreender a proposta de redação e aplicar conceitos das diversas áreas de conhecimento para desenvolver o tema, respeitando os limites estruturais do texto dissertativo-argumentativo.
3	Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.
4	Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da coesão textual.
5	Elaborar uma proposta de intervenção para o problema abordado, respeitando os direitos humanos e considerando sua viabilidade.

modelos em identificar pontos de melhoria nos textos. Informações detalhadas sobre os procedimentos adotados, os resultados obtidos e os exemplos de *feedbacks* podem ser consultadas no material suplementar do estudo, que visa apoiar a compreensão e a replicação das análises realizadas (ALBUQUERQUE, 2025).

4.0.4 Análise dos Resultados

A análise dos resultados envolveu a aplicação de métricas quantitativas para avaliar o desempenho dos modelos na correção automatizada. Inicialmente, foi calculada a média das notas atribuídas por cada modelo, permitindo identificar a tendência geral das avaliações. Em seguida, foi analisado o desvio padrão das notas, com o objetivo de verificar a consistência das correções e a estabilidade das pontuações geradas.

A taxa de acerto foi mensurada considerando a proximidade das avaliações automatizadas em relação ao padrão esperado no ENEM. Para isso, foi definida uma margem de variação tolerável em relação às notas atribuídas em correções manuais, permitindo verificar se os modelos treinados reduziram discrepâncias e aumentaram a precisão da avaliação. A análise também incluiu a verificação de tendências de viés, examinando se os modelos apresentavam padrões sistemáticos de atribuição de notas que pudessem comprometer a imparcialidade da correção.

Por fim, a análise qualitativa dos *feedbacks* gerados pelos modelos permitiu avaliar a profundidade e a relevância dos comentários oferecidos aos textos. Foram identificados padrões na estruturação dos *feedbacks*, observando-se aspectos como coerência na explicitação dos erros, clareza na sugestão de melhorias e adequação das recomendações pedagógicas.

5 RESULTADOS

Esta seção apresenta os resultados da avaliação comparativa entre o modelo GPT e as correções manuais, considerando sua capacidade de correção automatizada de redações do ENEM. A análise abrange a distribuição das notas (Seção 5.1), a consistência das avaliações (Seção 5.2) e a taxa de acerto em relação aos critérios oficiais do ENEM (Seção 5.3), permitindo uma reflexão aprofundada sobre as implicações do uso de GMLs para a avaliação de textos dissertativo-argumentativos. Detalhamentos adicionais sobre os dados analisados, exemplos de redações avaliadas e os relatórios gerados por cada modelo podem ser consultados no material suplementar do estudo (ALBUQUERQUE, 2025).

5.0.1 Distribuição das Notas

A Tabela 2 apresenta a comparação das notas reais e atribuídas pelo modelo GPT, destacando as diferenças entre a versão treinada e não-treinada.

Tabela 2 – Média e Comparação de Notas

Modelo	**C1**	**C2**	**C3**	**C4**	**C5**	**Nota Total**
GPT						
Nota real	154	145	123	160	132	715
GPT Não-treinado	127	125	108	115	107	582
GPT Treinado	145	132	135	125	115	652

A análise da distribuição das notas atribuídas pelos modelos revela diferenças substanciais entre as versões treinadas e não-treinadas. O GPT treinado apresentou uma média de 652 pontos, enquanto a versão não-treinada obteve 582 pontos, evidenciando um ganho significativo na capacidade do modelo de aderir aos critérios do ENEM após o ajuste fino.

Os valores mínimos e máximos das notas atribuídas pelos modelos refletem sua estabilidade e confiabilidade como ferramentas de avaliação. O GPT treinado demonstrou uma redução da discrepância entre as competências, aproximando-se das notas reais em cada critério. A versão não-treinada, por outro lado, subestimou significativamente a pontuação nas competências de desenvolvimento do tema (C2) e argumentação (C3), indicando dificuldades na identificação da estrutura textual e na diferenciação de redações de níveis distintos.

O desvio padrão das notas evidencia uma diferença significativa na estabilidade dos modelos. O GPT treinado reduziu a variabilidade das correções, aproximando suas avaliações das notas reais, o que reforça sua confiabilidade como ferramenta de avaliação automatizada. A análise comparativa reforça que o treinamento dos

modelos impacta diretamente a capacidade de alinhamento com os critérios avaliativos do ENEM, pois no caso do GPT, o ajuste fino resultou em avaliações mais precisas e estáveis.

5.0.2 Consistência das Avaliações

A consistência das avaliações é um fator determinante na viabilidade da automação da correção de redações, pois um sistema eficaz deve ser capaz de atribuir notas estáveis e coerentes entre diferentes competências e redações. O GPT treinado demonstrou uma redução significativa na discrepância entre as competências, evidenciando maior estabilidade na distribuição das notas. A análise revelou que, após o treinamento, o modelo apresentou menor variação entre as avaliações de coesão textual (C4) e estrutura argumentativa (C3), que historicamente apresentam maior subjetividade na correção humana. Esse comportamento sugere que o treinamento especializado permitiu ao modelo capturar melhor os padrões exigidos pelo ENEM, reduzindo flutuações excessivas entre diferentes critérios de avaliação. A versão não-treinada do GPT, por outro lado, apresentou uma maior dispersão nas notas entre as competências, refletindo uma dificuldade em manter um padrão estável de correção. Isso pode ser atribuído ao fato de que, sem o ajuste fino, o modelo não possuía um alinhamento adequado com os critérios específicos do ENEM, resultando em flutuações mais pronunciadas na atribuição de notas.

Outro aspecto relevante é a resposta dos modelos a redações de diferentes níveis de desempenho. O GPT treinado demonstrou maior capacidade de identificar erros gramaticais e estruturais, atribuindo notas mais próximas das avaliações humanas para redações com deficiências significativas. Já o GPT não-treinado frequentemente superestimou a pontuação de textos com problemas evidentes, demonstrando dificuldades em penalizar adequadamente erros linguísticos e argumentativos. Esse comportamento pode ser explicado pela ausência de refinamento específico para a tarefa de correção de redações do ENEM, o que impediu os modelos de capturar nuances da escrita dissertativo-argumentativa.

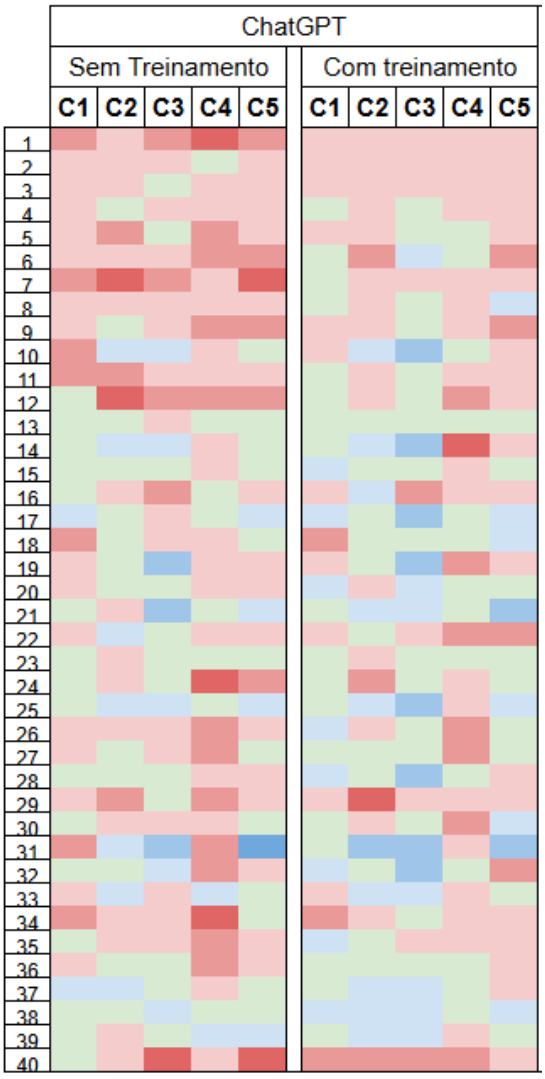
A avaliação da distribuição das notas entre as competências reforça que o treinamento do modelo impacta diretamente a sua consistência avaliativa. No caso do GPT, o ajuste fino resultou em uma melhor padronização da correção, reduzindo a variação excessiva entre critérios. As diferenças evidenciam, no entanto, que além do ajuste fino, abordagens complementares podem ser necessárias para garantir que os modelos consigam capturar adequadamente os critérios avaliativos do ENEM.

5.0.3 Taxa de Acerto em Relação aos Critérios do ENEM

A análise da calibração das avaliações do modelo GPT, considerando sua versão treinada e não-treinada, revela padrões distintos de erro e precisão. A Figura

2 apresenta um *heatmap* comparativo que evidencia a discrepância entre as notas atribuídas pelos modelos e as notas reais de referência. As colorações indicam a magnitude e a direção do erro: *verde* representa acertos exatos, enquanto tons de *vermelho* indicam subestimação e tons de *azul* indicam superestimação, com intensidades mais fortes indicando erros mais significativos.

Figura 2 – Taxa de acertos.



A distribuição dos erros evidencia que o GPT sem treinamento apresentou uma taxa média de 30% de acertos exatos, com melhor desempenho em C1 (37,5%) e C2 (35%) e menor precisão em C4 (20%), que avalia a coesão textual. Após o treinamento, houve uma melhoria substancial na precisão global do modelo, elevando a taxa de acertos para 47,5% em C1, 40% em C3 e 35% em C4, indicando um alinhamento mais próximo aos critérios do ENEM. Esse ganho foi mais evidente na competência C3 (organização e coerência argumentativa), que passou de 27,5% para 40%, sugerindo que o modelo treinado conseguiu capturar melhor a estruturação do texto dissertativo-argumentativo.

A análise da distribuição de erros por competência reforça que o ChatGPT treinado apresentou maior estabilidade e calibragem das notas, reduzindo significativamente flutuações extremas. Enquanto a versão não-treinada apresentava forte discrepância entre competências, sua versão treinada exibiu uma melhor distribuição dos erros, resultando em menos casos de superestimação e subestimação severas.

Outro aspecto relevante é o comportamento dos modelos frente a redações de baixo desempenho. O ChatGPT treinado demonstrou maior capacidade de penalizar adequadamente textos com falhas estruturais, diferenciando com mais precisão os níveis de qualidade textual. Enquanto na versão não-treinada algumas redações mal formuladas recebiam notas acima do esperado, o modelo treinado ajustou esse comportamento, reduzindo a taxa de superestimação.

6 LIMITAÇÕES E AMEAÇAS À VALIDADE

Esta pesquisa apresenta algumas ameaças à validade, conforme a classificação de Wohlin et al. (WOHLIN *et al.*, 2012), e ações para mitigar seus efeitos, garantindo maior rigor nos resultados.

Validade Interna. A calibração dos modelos representa um risco, pois a avaliação automatizada depende da correta interpretação dos critérios do ENEM. Possíveis vieses nos conjuntos de treinamento podem ter influenciado as avaliações. Para mitigar esse problema, utilizamos um conjunto representativo de redações, abrangendo diferentes níveis de qualidade. Outra ameaça refere-se à distribuição dos erros. O GPT não-treinado apresentou forte tendência à subestimação das notas, enquanto o GPT treinado mostrou maior alinhamento com as avaliações humanas. Para minimizar esse viés, foram utilizadas múltiplas competências como referência, reduzindo variações individuais na atribuição das notas. *Validade de Construção*. Como os critérios de correção do ENEM são aplicados por humanos e contêm elementos subjetivos, há o risco dos GMLs aprenderem padrões inconsistentes. Para reduzir essa ameaça, ambos modelos foram treinados com exemplos representativos, e a análise combinou métricas quantitativas e qualitativas para avaliar a aderência aos critérios oficiais. Outra limitação está na métrica de acerto exato, que pode não refletir pequenas variações aceitáveis na nota. Para evitar interpretações equivocadas, complementamos a análise observando a distribuição dos erros e a consistência das avaliações por competência.

Validade Externa. A generalização dos resultados para outros contextos avaliativos é uma possível limitação. Embora o estudo tenha seguido os critérios do ENEM, seus achados podem não se aplicar a outros exames com escalas de avaliação distintas. Para mitigar essa ameaça, detalhamos a metodologia adotada, permitindo que futuros estudos avaliem sua replicabilidade. Outra ameaça diz respeito ao contexto temporal. Os GMLs evoluem rapidamente, e novas versões do GPT podem apresentar desempenhos diferentes. Para minimizar esse impacto, os dados foram coletados em um período representativo, considerando as limitações das versões utilizadas. *Validade de Conclusão*. O tamanho da amostra pode ser uma limitação. Embora tenhamos analisado 40 redações, uma amostra maior poderia oferecer resultados mais robustos. Para mitigar essa questão, além da taxa de acertos exatos, examinamos a distribuição dos erros e a consistência das avaliações entre as competências. Outro risco envolve possíveis vieses na atribuição das notas de referência. Como a correção humana pode conter variações subjetivas, os critérios do ENEM foram rigorosamente seguidos na construção da base de avaliação, garantindo maior precisão na calibração dos modelos.

7 CONSIDERAÇÕES FINAIS

Este estudo analisou a viabilidade do uso de GMLs na correção automatizada de redações do ENEM, comparando versões treinadas e não-treinadas do modelo GPT. Os resultados evidenciaram que o GPT treinado apresentou maior aderência às notas reais, com ganhos expressivos em precisão e estabilidade, especialmente na competência C3 (organização e coerência textual), cuja taxa de acerto aumentou de 27,5% para 40% após o ajuste fino.

A principal contribuição desta pesquisa está em demonstrar o impacto positivo do treinamento supervisionado na calibração das avaliações automatizadas, indicando que GMLs como o GPT, quando adequadamente ajustados, podem se tornar ferramentas eficazes e confiáveis no contexto da educação. Os resultados também reforçam a importância de abordagens híbridas — combinando inteligência artificial com a revisão humana — como estratégia promissora para garantir mais equidade, eficiência e padronização na correção de redações em larga escala.

Como trabalhos futuros, busca-se aprimorar a taxa de acertos nas avaliações quantitativas por meio de novos treinamentos dos modelos e estratégias de calibração, reduzindo discrepâncias na atribuição de notas. Além disso, será realizada uma avaliação qualitativa dos *feedbacks* gerados pelos GMLs, com a participação de professores especialistas, analisando sua clareza, coerência e aderência aos critérios do ENEM. Por fim, pretende-se desenvolver uma ferramenta completa e customizada para professores e alunos, ampliando seu uso tanto para correção automatizada quanto para suporte pedagógico na escrita.

REFERÊNCIAS

ALBUQUERQUE, D. W. **Material Suplementar - Avaliação Automatizada de Redações do ENEM: Um Estudo Comparativo entre Grandes Modelos de Linguagem**. figshare, 2025. Dataset.https://doi.org/10.6084/m9.figshare.29361035. Disponível em: <https://doi.org/10.6084/m9.figshare.29361035>. Citado 3 vezes nas páginas 18, 20 e 21.

ALTAMIMI, A. B. Effectiveness of chatgpt in essay autograding. In: IEEE. **2023 International Conference on Computing, Electronics & Communications Engineering (iCCECE)**. [S.l.], 2023. p. 102–106. Citado na página 16.

ANDRADE, E. **Material Suplementar - Avaliação Automatizada de Redações do ENEM: Um Estudo Comparativo entre Grandes Modelos de Linguagem**. Google Drive, 2025. Disponível em: <https://drive.google.com/drive/folders/1ftl_ZIXYsIAFWb4e5NBekkxWAPUeoSpO?usp=sharing>. Citado na página 19.

BUI, N. M.; BARROT, J. S. Chatgpt as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. **Education and Information Technologies**, Springer, p. 1–18, 2024. Citado na página 16.

CARNEIRO, J. R. S. *et al.* O uso do google sala de aula na educação básica: uma perspectiva pedagógica convergente à educação contextualizada no ifrn. Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte, 2018. Citado na página 14.

FITRIA, T. N. Artificial intelligence (ai) technology in openai chatgpt application: A review of chatgpt in writing english essay. In: **ELT Forum: Journal of English Language Teaching**. [S.l.: s.n.], 2023. v. 12, n. 1, p. 44–58. Citado na página 16.

GUIMARÃES, U. A. *et al.* As mídias digitais no campo educacional: Um olhar pelas aplicações do chat gpt na educação. **RECIMA21-Revista Científica Multidisciplinar-ISSN 2675-6218**, v. 4, n. 7, p. e473556–e473556, 2023. Citado na página 13.

GUZMAN, J. H. E. *et al.* Information and communication technologies (ict) in the processes of distribution and use of knowledge in higher education institutions (heis). **Procedia Computer Science**, Elsevier, v. 198, p. 644–649, 2022. Citado na página 14.

INEP. **Cartilha do Participante – Redação no Enem 2023. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)**. 2023. Acessado em: 03 mar. 2025. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/noticias/enem/cartilha-do-participante-redacao-no-enem-2023>. Citado na página 14.

JÚNIOR, C. R.; SPALENZA, M. A.; OLIVEIRA, E. de. Proposta de um sistema de avaliação automática de redações do enem utilizando técnicas de aprendizagem de máquina e processamento de linguagem natural. **Anais do Computer on the Beach**, v. 8, p. 474–483, 2017. Citado na página 16.

LEE, A. **What Are Large Language Models Used For?** 2023. <https://blogs.nvidia.com/blog/what-are-large-language-models-used-for/>. Acesso em: 03 dez. 2023. Citado 2 vezes nas páginas 13 e 14.

LIMA, J. Como o chatgpt afeta a educação e o desenvolvimento universitário. **The Trends Hub**, n. 3, 2023. Citado 2 vezes nas páginas 13 e 14.

LIMA, T. B. de *et al.* Avaliaçao automática de redaçao: Uma revisao sistemática. **Revista Brasileira de Informática na Educação**, v. 31, p. 205–221, 2023. Citado na página 16.

MONTEIRO, J. C. da S. Assistente chatgpt na educação: Possibilidades e desafios. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, v. 9, n. 6, p. 2899–2906, 2023. Citado 2 vezes nas páginas 13 e 14.

NETO, S. S. C. *et al.* Avaliação automática de redações na língua portuguesa baseada na coleta de atributos e aprendizagem de máquina. In: SBC. **Simpósio Brasileiro de Informática na Educação (SBIE)**. [S.l.], 2020. p. 1162–1171. Citado na página 16.

VARAS, J. *et al.* Innovations in surgical training: exploring the role of artificial intelligence and large language models (llm). **Revista do Colégio Brasileiro de Cirurgiões**, SciELO Brasil, v. 50, p. e20233605, 2023. Citado na página 14.

WOHLIN, C. *et al.* **Experimentation in software engineering**. [S.l.]: Springer Science & Business Media, 2012. Citado na página 25.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Campina Grande - Código INEP: 25137409
R. Tranqüilino Coelho Lemos, 671, Dinamérica, CEP 58432-300, Campina Grande (PB)
CNPJ: 10.783.898/0003-37 - Telefone: (83) 2102.6200

Documento Digitalizado Ostensivo (Público)

Versão Final do TCC

Assunto:	Versão Final do TCC
Assinado por:	Eliel Souza
Tipo do Documento:	Anexo
Situação:	Finalizado
Nível de Acesso:	Ostensivo (Público)
Tipo do Conferência:	Cópia Simples

Documento assinado eletronicamente por:

- **Eliel Alexandre Andrade Souza**, **ALUNO (201811250037) DE BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO - CAMPINA GRANDE**, em 01/09/2025 12:20:08.

Este documento foi armazenado no SUAP em 01/09/2025. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1592721
Código de Autenticação: 1161a20269