SOL Preprints, 2026 • doi: 10.5753/sol.2026.0003

Artigo de Pesquisa/Research Paper

Geração de Feedbacks Educacionais Personalizados por IA Generativa: Validação do ChatGPT no Contexto do Ensino Médio Brasileiro

Generation of Personalized Educational Feedback by Generative AI: Validation of ChatGPT in the Context of Brazilian High School Education

Luiz Antônio Oliveira de Melo
Universidade Federal Rural de Pernambuco
luiz.oliveiramelo@ufrpe.br

Taciana Pontual Falcão
Universidade Federal Rural de Pernambuco
taciana.pontual@ufrpe.br

Resumo. Este trabalho investiga a eficácia do uso de Inteligência Artificial Generativa, especificamente o ChatGPT, na geração de feedbacks para redações do ensino médio brasileiro. Utilizando o corpus Essay-BR, o estudo comparou avaliações realizadas pela IA com correções humanas, aplicando métricas de correlação e erro (MAE/RMSE), além de uma análise qualitativa baseada em autoavaliação do modelo. Os resultados quantitativos indicaram uma correlação moderada e uma tendência de superavaliação das notas pela IA. Por outro lado, a análise qualitativa demonstrou que o ChatGPT é capaz de gerar feedbacks estruturalmente coerentes, polidos e com caráter motivador. Conclui-se que a ferramenta, no estágio atual, apresenta maior potencial como instrumento de suporte educacional do que como avaliadora independente.

Abstract. This work investigates the effectiveness of using Generative Artificial Intelligence, specifically ChatGPT, to generate feedback for Brazilian high school essays. Using the Essay-BR corpus, the study compared AI evaluations with human corrections, applying correlation and error metrics (MAE/RMSE), alongside a qualitative analysis based on the model's self-assessment. Quantitative results indicated a moderate correlation and a tendency for the AI to overestimate scores. On the other hand, the qualitative analysis showed that ChatGPT is capable of generating structurally coherent, polished, and motivating feedback. In conclusion, the tool, in its current stage, presents greater potential as an educational support instrument rather than as an independent evaluator.

Palavras-chave: Feedback Personalizado, IA Generativa, ChatGPT

Keywords: Personalized Feedback, Generative AI, ChatGPT

Recebido/Received: 26 Jan 2026 • **Publicado/Published:** 20 Feb 2026

1 Introdução

O feedback é um dos pilares fundamentais para o desenvolvimento positivo do aluno no processo de ensino-aprendizagem, potencializando a autorregulação e causando reflexão sobre seus pontos fortes e fracos [Costa Júnior *et al*., 2023]. O ensino personalizado busca enxergar e trabalhar as individualidades de cada aluno [da Silva Neta and Capuchinho, 2017], colocando-o no centro de todo o processo e rompendo com o sistema tradicional onde o professor é o foco principal.

Contudo, essa estratégia também apresenta dificuldades principalmente quando pensamos na atuação em instituições de educação básica. Por si só as atividades realizadas pelos professores vão além dos momentos em sala de aula, adicionando o fato de muitos professores lecionarem em mais de um local e mais de um turno, nos diferentes níveis de ensino, infantil, fundamental e médio. O estudo de [Guerreiro *et al*., 2016] realizou uma pesquisa com 978 professores da educação básica, onde 72,9% tinham mais de um local de trabalho e 80,8% lecionavam em mais de um turno, resultando em uma carga de trabalho altíssima. Vale também ressaltar que 29,4% desses professores apontaram quantidade alta de alunos por sala de aula.

Pensando nesses dados, é notável a problemática de personalizar o ensino para cada um desses estudantes. Alguns professores procuram modos de tornar isto possível, fazendo uso de tecnologia digitais a fim de diminuir sua carga horária e abrir margem para melhorias do processo de ensino-aprendizagem [da Silva Neta and Capuchinho, 2017].

Com a recente alta da utilização da Inteligência Artificial (IA) tem surgido uma variedade de estudos que buscam sua utilização para o ambiente educacional. Como resultado, ferramentas para auxílio do ensino de programação, matemática e línguas estrangeiras foram desenvolvidas, ou aprimoradas, para utilizar IA. São exemplos: Avatutor, Cognitive Tutor e Duolingo [Giraffa and Kohls-Santos, 2023]. Outro exemplo é o Pounce, uma plataforma de análise preditiva que analisa dados dos alunos e fornece alertas de alunos predispostos a abandonar a universidade [Picão *et al*., 2023].

Essas ferramentas citadas geram feedbacks personalizados das atividades, sendo essa uma das principais vantagens da IA para educação, a personalização do ensino e feedback

imediato [Di Santo, 2023]. Por outro lado, o uso da IA traz consigo problemas como respostas com vieses, erros de análise e questões relacionadas à privacidade dos dados dos estudantes [Picão *et al*., 2023].

O problema central desta pesquisa reside na lacuna entre a promessa tecnológica e a validação pedagógica: embora ferramentas de Inteligência Artificial Generativa (IAg), como o ChatGPT, demonstrem capacidade para gerar respostas contextualizadas, emerge a seguinte questão de pesquisa:

"A IAg pode ser operacionalizada para gerar feedbacks educacionais personalizados e **válidos**?"

É importante salientar, desde já, que a validação proposta neste estudo se desdobra em duas dimensões distintas de análise. A primeira é de natureza quantitativa/somativa, onde investiga-se a precisão da IA em atribuir notas numéricas comparáveis às de avaliadores humanos. A segunda, que é o foco central da aplicação pedagógica, é a dimensão qualitativa/formativa, que analisa a estrutura e a utilidade do texto de feedback gerado para orientar a reescrita do aluno. Essa distinção é fundamental, pois permite avaliar a ferramenta não apenas como um "corretor automático", mas como um potencial assistente de tutoria. Adicionalmente, aplica-se uma análise de sentimentos sobre esta avaliação qualitativa, com o objetivo de verificar se a análise textual produzida pelo modelo mantém um tom de neutralidade e segurança, isento de toxicidade ou viés emocional desproporcional.

Assim, este trabalho tem como objetivo geral investigar a eficácia do ChatGPT na geração de feedbacks para redações de nível ensino médio. A escolha por esta ferramenta específica justifica-se pela sua ampla difusão no contexto educacional.

Os objetivos específicos são:

1. Comparar a concordância entre as notas atribuídas pela IAg e a avaliação humana, utilizando métricas estatísticas de correlação e erro;
2. Avaliar a qualidade pedagógica dos feedbacks gerados, utilizando critérios considerados fundamentais por estudantes e professores para um feedback eficaz;
3. Analisar a consistência dos resultados (notas e feedbacks) gerados pelo modelo diante de múltiplas execuções (divergência);
4. Verificar a adequação do tom e a segurança das análises qualitativas (autoavaliações) geradas pelo modelo, aplicando técnicas de análise de sentimentos para identificar viés emocional ou toxicidade nesta etapa de validação.

Sua relevância assenta-se em validar uma solução replicável para redução da carga de trabalho docente em correções de redações.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta o referencial teórico que fundamenta a pesquisa; a Seção 3 cita os trabalhos relacionados; a Seção 4 descreve a metodologia utilizada para coleta, geração e validação dos dados; a Seção 5 detalha a análise dos resultados quantitativos e qualitativos obtidos; e a Seção 6 apresenta as conclusões e sugestões para trabalhos futuros.

2 Referencial Teórico

2.1 Ensino Personalizado e Desafios Docentes

O ensino personalizado surge com uma abordagem pedagógica para adaptar as experiências do processo de ensino-aprendizagem para a individualidade dos estudantes. A personalização consiste em ajustar o conteúdo, o ritmo e as estratégias, tornando a educação mais engajadora e eficaz, chegando ao feedback personalizado que envolve a prática de reagir ao envolvimento individualizado do discente em todo processo, criando aberturas para melhor desenvolvimento educacional [Silva *et al*., 2025].

Porém, todo este processo exige uma carga ainda maior dos docentes, considerando o número de estudantes para os quais ele leciona. Desse modo procuram-se estratégias que permitam aos discentes terem o melhor para se desenvolverem sem que cause ainda mais sobrecargas de trabalho aos professores [Cunto *et al*., 2022]. Neste sentido, observa-se o uso da IA com algumas vantagens quando usada neste processo de ensino-aprendizagem. Como, por exemplo, a possibilidade de feedback imediato e acesso a conteúdos de qualidade [Picão *et al*., 2023], para os estudantes, além de auxiliar o professor na construção das aulas [Di Santo, 2023].

2.2 Evolução da IA na Educação: Dos Sistemas Tutores à IA Generativa

A aplicação da Inteligência Artificial na educação não é um fenômeno recente, mas tem passado por transformações paradigmáticas. Historicamente, a área consolidou-se através dos Sistemas Tutores Inteligentes (STI), que são sistemas arquitetados com módulos explícitos para representar o domínio do conhecimento, o modelo do aluno e o modelo pedagógico [Giraffa and Kohls-Santos, 2023]. Esses sistemas tradicionais operam baseados em regras rígidas e caminhos de aprendizagem pré-definidos, sendo eficazes para domínios específicos, mas limitados em flexibilidade linguística.

Recentemente, houve uma ruptura com a emergência da Inteligência Artificial Generativa (IAg). Diferente dos STIs, a IAg opera através de modelos probabilísticos de linguagem (LLMs) treinados em vastos volumes de dados textuais, permitindo interações em linguagem natural com fluidez inédita [Giraffa and Kohls-Santos, 2023]. No contexto atual, embora o ChatGPT seja a ferramenta mais difundida, o ecossistema expandiu-se com modelos como Gemini (Google), Claude (Anthropic) e Llama (Meta). Essas tecnologias, somadas a sistemas de Learning Analytics, compõem o novo cenário de personalização, oferecendo oportunidades para adaptar conteúdo às necessidades individuais, apesar dos desafios de infraestrutura no contexto brasileiro [Silva *et al*., 2025].

2.3 O ChatGPT e Modelos de Linguagem

Com a IA, pode-se desenvolver sistemas capazes de realizar tarefas que envolvem habilidades como resolução de problemas e tomadas de decisões. O uso dessas tecnologias vem desempenhando um papel cada vez mais importante, em diferentes contextos [Di Santo, 2023]. O ChatGPT (*Generative Pre-trained Transformer*) é um exemplo proeminente de Modelo de Linguagem de Grande Escala (LLM). É uma ferramenta baseada em um modelo de IAg treinado a partir de

quantidades massivas de dados, capaz de gerar conteúdos semelhantes à escrita humana [Adelina Moura, 2023], sendo capaz de responder perguntas, fornecer informações e até simular um ambiente de conversação, tudo isso mediante comandos/instruções fornecidas pelo usuário.

2.4 Literacia de Prompts

A literacia de prompts é a habilidade de formular essas instruções, tornando-as mais precisas para a IAg, visando respostas relevantes [Adelina Moura, 2023]. Seu domínio é essencial para obter um feedback automatizado de qualidade, pois a estrutura de interações deve possuir um formato simples que previna ambiguidades, detalhar suficientemente as necessidades e direcionar o fluxo de interação com a ferramenta de IAg.

3 Trabalhos Relacionados

A literatura sobre avaliação automática de escrita e geração de feedbacks tem evoluído significativamente, transitando de métodos baseados em contagem de características superficiais para o uso de LLMs. Esta seção destaca estudos fundamentais que oferecem tanto a base de dados para o contexto brasileiro quanto metodologias de validação de IAg.

3.1 Recursos e Avaliação Automática em Português

No cenário da língua portuguesa, o desenvolvimento de sistemas de Avaliação Automática de Redações (AES - *Automated Essay Scoring*) enfrentou historicamente a escassez de corpora anotados publicamente. Para preencher essa lacuna, Marinho *et al*. [2021] propuseram o **Essay-BR**, o primeiro corpus público de grande escala contendo 4.570 redações argumentativas escritas por estudantes brasileiros.

Diferentemente das abordagens baseadas em LLMs, Marinho *et al*. [2021] estabeleceram linhas de base (*baselines*) utilizando métodos de aprendizado de máquina tradicionais. Os autores empregaram técnicas baseadas em extração de características manuais (*hand-crafted features*), como a contagem de erros gramaticais, verbos e marcadores discursivos, alimentando modelos de regressão linear e *Gradient Boosting* para predizer as notas. Embora fundamentais para a área, esses métodos focam predominantemente na atribuição de notas (holísticas ou por competência) e dependem de engenharia de características explícita, limitando-se na capacidade de prover explicações semânticas profundas ou feedback formativo detalhado ao aluno, foco da presente pesquisa.

3.2 Feedback Gerado por IA Generativa

Com a emergência dos modelos de IAg, o foco das pesquisas expandiu-se da simples atribuição de notas para a geração de feedback qualitativo. Um estudo recente de Zhu *et al*. [2024] investigou o potencial do ChatGPT na geração de feedback para escrita em inglês, *English as a Foreign Language* (EFL), comparando-o diretamente com correções realizadas por professores humanos. Os autores categorizaram a atuação da IA em três níveis: modificação primária (gramática e pontuação), refinamento intermediário (estilo e vocabulário) e produção de alto nível (resumos e lógica). Os resultados indicaram que o ChatGPT supera métodos tradicionais ao oferecer correções mais naturais e fluidas, reduzindo o ”tiques”de escrita não nativa, além de demonstrar maior objetividade e consistência que os avaliadores humanos, que tendem a ser influenciados por subjetividades pessoais

Entretanto, Zhu *et al*. [2024] concluem que a IA não substitui o docente, pois ainda carece de ”inteligência emocional”e capacidade crítica profunda para nuances textuais complexas. Os autores sugerem uma abordagem de ”via dupla”(*dual track*), onde a IA atua no suporte imediato e correção estrutural, enquanto o professor foca no desenvolvimento estratégico do aluno. Este trabalho corrobora essa visão, aplicando uma metodologia de validação similar ao contexto das redações do ENEM disponibilizadas pelo corpus Essay-BR.

4 Metodologia

O presente trabalho se propõe a investigar a eficácia do ChatGPT na geração de feedbacks para redações de nível ensino médio. O método para atingir este objetivo é composto de três etapas, apresentadas na Figura 1 e descritas a seguir:

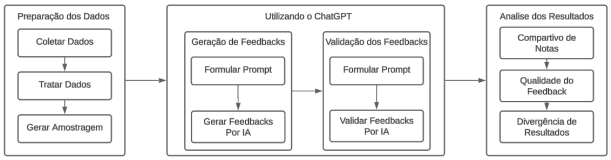

Figura 1. Etapas da metodologia aplicada ao estudo.

4.1 Preparação dos Dados

4.1.1 Coleta dos dados

A base de dados utilizada neste estudo é o Corpus **Essay-BR**, base pública composta por redações produzidas por estudantes do ensino médio e corrigidas por especialistas seguindo os critérios avaliativos do ENEM vigentes em 2021. Este conjunto de dados foi constituído especificamente com o propósito de servir para o desenvolvimento de novos métodos de correção [Marinho *et al*., 2021]. A base Essay-BR se encontra disponível no GitHub: https://github.com/rafaelanchieta/essay.

4.1.2 Tratamento de Dados e Geração de Amostra

Após a coleta, foi realizado o pré-processamento da base a fim de adequá-la ao objetivo deste estudo. Inicialmente, o corpus **Essay-BR** foi carregado em ambiente Google Colab, resultando em um conjunto de 4.570 redações acompanhadas de informações adicionais, tais como título, nota final, prompt e competências avaliadas.

Na etapa de inspeção inicial, identificou-se que 684 redações não possuíam título definido. Essas instâncias foram descartadas, uma vez que a ausência de título comprometeria análises posteriores. Em seguida, realizou-se uma filtragem com expressões regulares para remover títulos inválidos, tais como sequências de caracteres não representativos (pontos, números isolados, letras soltas) e a indicação “sem título”. Esse procedimento garantiu que apenas redações com títulos consistentes fossem mantidas no corpus.

Além disso, a coluna **prompt** foi excluída, pois não era relevante para os objetivos do estudo. Dessa forma, a base

final foi reduzida a quatro atributos principais: título, redação (texto), nota e competências.

No que se refere à variável **nota**, observou-se que os valores variam de 0 a 1000, conforme os critérios oficiais de avaliação do ENEM, distríbuidas de acordo com os critérios (competências) definidos. Para possibilitar a seleção de um conjunto representativo, as notas foram normalizadas em **faixas de desempenho** utilizando a técnica de quantis. Foram estabelecidas cinco categorias: *muito baixa*, *baixa*, *média*, *alta* e *muito alta*. Essa classificação permitiu distribuir as redações de acordo com diferentes níveis de desempenho.

Por fim, foram selecionadas **cem redações** para compor a amostra final, sendo vinte provenientes de cada faixa de desempenho. Essa escolha buscou garantir equilíbrio entre diferentes níveis de qualidade textual, de modo a favorecer a avaliação dos feedbacks gerados pelo modelo de linguagem. Assim, considerando-se as limitações de custo/tempo de processamento da API paga, a abordagem mantém o equilíbrio entre as 5 faixas de nota (20 redações cada) para garantir validade estatística.

4.2 Geração de Feedbacks

4.2.1 Formulação de Prompt

Para definir o prompt a ser enviado juntamente com as redações, foi usado o modelo *RACE* [Adelina Moura, 2023]. A construção do prompt seguiu rigorosamente a estrutura deste framework para minimizar alucinações e padronizar a saída:

- *Role* - A persona 'professor de língua portuguesa' foi ativada para modular o tom e o vocabulário da resposta, aproximando-a de uma correção pedagógica humana.
- *Action* - O comando explícito para "avaliar redações e fornecer feedbacks" delimitou a tarefa, impedindo que o modelo apenas resumisse ou reescrevesse o texto.
- *Context* - Foram inseridos explicitamente os 5 critérios oficiais de correção do ENEM para garantir que o modelo não utilizasse critérios genéricos de escrita criativa, mas sim a régua técnica do exame.
- *Expectation* - A saída foi restringida a um formato estruturado (JSON/Chave-Valor) para permitir a extração automatizada dos dados, separando a nota numérica do texto qualitativo.

Seguindo este modelo, foi elaborado o seguinte prompt: "Assuma o papel de um professor de língua portuguesa do ensino médio (*Role*), avalie redações e faça feedbacks para cada uma delas (*Action*). Utilize os critérios do ENEM para fazer a avaliação, são eles: Respeito à norma formal da escrita portuguesa; Conformidade com o gênero textual argumentativo e o tema proposto (enunciado), para desenvolver um texto utilizando conhecimentos de diferentes áreas; Seleção, relação, organização e interpretação de dados e argumentos em defesa de um ponto de vista; Utilização de estruturas linguísticas argumentativas; e Elaboração de uma proposta para solucionar o problema em questão (*Context*). Espera-se as notas por competência no intervalo de 0 a 200, a nota total esteja no intervalo entre 0 e 1000, seguindo o ENEM e que o feedback seja capaz de contribuir com o processo de ensino-aprendizagem do aluno, a resposta deve obrigatoriamente seguir a estrutura: title: [Título da redação original, sem alterações], score-GPT: [nota total entre 0 e 1000], competence-GPT: [C1: [0–200], C2: [0–200], C3: [0–200], C4: [0–200], C5: [0–200]], feedback-GPT: [feedback gerado] (*Expectation*)."

Ressalta-se que a solicitação explícita da nota total (Score-GPT) no prompt, juntamente com as competências, teve como objetivo reforçar o contexto da escala global do ENEM (0-1000) para o modelo, garantindo que a avaliação considerasse a coerência entre as partes e o todo.

4.2.2 Produção dos Feedbacks

Para gerar os feedbacks, no ambiente do *Google Colab* onde as redações foram carregadas previamente, foi utilizada a API da OpenAI para o ChatGPT. A fim de garantir uma versão disponível de forma gratuita, sendo mais acessível e estável, foi usado o modelo do ChatGPT 3.5-turbo, inciando uma nova interação. Para realizar tal operação, foi necessário ter uma conta de acesso à plataforma da OpenAI, gerar uma API Token e realizar o cadastro do metódo de cobrança para uso da API. A versão ChatGPT 3.5-turbo via web é gratuita, a cobrança existe apenas por uso da API no ambiente de desenvolvimento.

O prompt previamente definido foi estrurado em três camadas:

1. o *system prompt*, que orienta o modelo a assumir o papel de um professor do ensino médio - define o *Role*;
2. o *user prompt*, onde são especificados o que deve feito explicitando os detalhes envolvidos - define *Action*, *Context* e *Expectation*.
3. a inserção programática das redações, permitindo a avaliação em série - define o conteúdo que será explorado.

O resultado retornado pelo ChatGPT foi salvo em um arquivo texto para manter a primeira resposta gerada, dado que se refeita pode gerar resultados diferentes. Também foi preciso tratar o retorno da interação, que vem como *string* pura. Assim, para organizar os resultados retornados pelo modelo, foi estruturado um *DataFrame* contendo quatro colunas principais: **title**, que registra o identificador da redação; **score_GPT**, que armazena a nota global atribuída pelo modelo; **competence_GPT**, composta por um vetor numérico representando as cinco competências avaliadas; e **feedback_GPT**, que contém o texto descritivo do retorno gerado pelo ChatGPT, posteriormente foi agregado com as informações do DataFrame original da base Essay-BR. Essa etapa de tabulação é essencial para permitir análises posteriores, filtragens e integração com métodos estatísticos. A fim de manter os dados já estruturados também foi salvo o *DataFrame* em CSV.

4.3 Validação dos Feedbacks

Para a avaliação qualitativa da estrutura dos feedbacks gerados, adotou-se a abordagem metodológica conhecida como *LLM-as-a-Judge* (LLM como Juiz). Esta técnica consiste em utilizar Modelos de Linguagem de Grande Escala para avaliar a qualidade de textos gerados por outros modelos (ou por si mesmos) - uma prática que tem demonstrado alta correlação com julgamentos humanos em tarefas de geração de linguagem natural [Liu *et al.*, 2023].

A escolha por esta autoavaliação guiada justifica-se pela escalabilidade. De modo semelhante ao processo anterior,

foi estruturado um *prompt* e iniciada uma nova interação com a ferramenta para realizar essa análise.

4.3.1 Formulação do Prompt

O modelo RACE foi novamente aplicado para estruturar a validação, adaptando o contexto para a tarefa de validação.

- *Role* - A IA foi instruída novamente como a persona de 'professor', garantindo que a validação da qualidade do feedback seguisse uma perspectiva pedagógica e não apenas linguística.
- *Action* - A tarefa consistiu em analisar os pares "Redação-Feedback" para julgar se o retorno fornecido seria útil para o aluno.
- *Context* - Os critérios de qualidade foram fundamentados no trabalho de [Pontual Falcão *et al*., 2023], definindo explicitamente quatro dimensões essenciais: explicação do erro, indicação de lacunas (o que falta), sugestões de melhoria e aspectos positivos.
- *Expectation* - Exigiu-se uma avaliação quantitativa em escala Likert (0 a 5) para cada dimensão, além de uma análise qualitativa, permitindo a mensuração estatística da qualidade pedagógica percebida.

Desse modo, o seguinte *prompt* foi gerado: "Assuma o papel de um professor de língua portuguesa do ensino médio (*Role*), avalie as seguintes redações e os feedbacks apresentados (*Action*). Considere os seguintes critérios para determinar a qualidade de cada feedback: explicar o que está errado, apresentar o que está faltando, sugerir pontos a serem melhorados e destacar aspectos positivos para motivar o estudante (*Context*). Cada critério deve ser avaliado em uma escala de 0 (zero) a 5 (cinco), sendo 0 muito ruim e 5 muito bom. Além disso, forneça uma breve análise qualitativa de cada feedback, a resposta deve obrigatoriamente seguir a seguinte estrutura: title: [titulo da redação original, sem alterações], explicação do que está errado: [intervalo de 0 a 5], apresentar o que está faltando [intervalo de 0 a 5], pontos a serem melhorados: [intervalo de 0 a 5], aspectos positivos: [intervalo de 0 a 5] (*Expectation*)."

4.3.2 Validação dos Feedbacks

A validação dos feedbacks foi feita de maneira bem semelhante à que foi feita na etapa de geração dos feedbacks, sendo a única mudança o prompt. A estrutura segue em três camadas, porém desta vez além das redações, os feedbacks anteriormente gerados foram adicionados ao conteúdo a ser explorado.

A resposta do modelo foi salva em arquivo texto para manter a imutabilidade da interação. Para estruturar os resultados obtidos na etapa de validação, o retorno também foi convertido em um *DataFrame* contendo as variáveis essenciais para análise. As colunas registram: **Título**, correspondente ao identificador da redação original, foi traduzido após retorno do modelo; **Explicação_Errado**, **Apresentar_Faltando**, **Pontos_Melhorar** e **Aspectos_Positivos**, que armazenam as pontuações numéricas atribuídas pelo modelo para cada um dos critérios avaliados; além da coluna **Análise_Qualitativa**, que descreve de maneira qualitativa o feedback produzido pelo ChatGPT. Ao final, o *DataFrame* foi exportado em CSV, a fim de facilitar a reutilização em etapas posteriores.

5 Análise dos Resultados

Nesta seção, são apresentados os resultados gerados pelo ChatGPT nas duas interações e as análises com base nas notas e na qualidade dos feedbacks.

Para a validação quantitativa, foram selecionadas métricas amplamente adotadas na literatura de Avaliação Automática de Redações (*Automated Essay Scoring* - AES). A Correlação de Pearson (r) foi utilizada para medir a força da associação linear entre as notas atribuídas pelo ChatGPT e pelos avaliadores humanos, indicando a consistência da predição em relação ao padrão-ouro. Complementarmente, adotou-se o Erro Absoluto Médio (MAE) e a Raiz do Erro Quadrático Médio (RMSE) para quantificar a magnitude da divergência nas notas. Essas métricas são o padrão de referência para comparabilidade com o estado da arte em língua portuguesa, conforme estabelecido por estudos anteriores (AMORIM; VELOSO, 2017 *apud* [Marinho *et al*., 2021]).

5.1 Análise das Notas

A avaliação inicial buscou comparar as notas atribuídas pelo ChatGPT com aquelas concedidas pelos avaliadores humanos, tanto para a nota total quanto para cada uma das cinco competências exigidas pelo ENEM. Para isso, foram calculadas medidas de correlação, erro e viés.

5.1.1 Correlação entre ChatGPT e avaliadores humanos

A Tabela 1 apresenta os coeficientes de correlação de Pearson entre as duas avaliações. Os resultados indicam correlações **baixas a moderadas** nas competências e uma correlação **moderada** para a nota total.

Tabela 1. Correlação entre competências geradas por humanos e ChatGPT

Sigla	Descrição da Competência	Correlação
C1	Domínio da escrita formal	0,34
C2	Compreensão do tema e gênero	0,31
C3	Seleção e organização de argumentos	0,37
C4	Mecanismos linguísticos (Coesão)	0,36
C5	Proposta de intervenção	0,23
Nota Total		**0,45**

Esses valores mostram que, embora exista relação positiva entre as avaliações, o alinhamento ainda é limitado. A maior correlação ocorre na nota global (0,45), sugerindo que o modelo captura parte da progressão geral de qualidade das redações, mas apresenta dificuldades em se alinhar às especificidades de cada competência.

A menor correlação ocorre na Competência 5 (0,23), o que confirma a dificuldade do modelo em avaliar critérios ligados à proposta de intervenção.

5.1.2 Erros médios (MAE e RMSE)

A Tabela 2 apresenta os valores de Erro Médio Absoluto (MAE) e Raiz do Erro Quadrático Médio (RMSE) entre as avaliações humanas e as realizadas pelo modelo.

Os resultados obtidos demonstram que o modelo ChatGPT alcançou um desempenho competitivo em relação a métodos tradicionais de aprendizado de máquina treinados

Tabela 2. Métricas de erro (MAE e RMSE) do modelo por competência

Competência	MAE	RMSE
C1	29,68	37,88
C2	34,60	44,58
C3	34,73	44,03
C4	34,17	43,82
C5	48,76	62,01
Score Total	140,34	169,96

especificamente para esta tarefa. O RMSE total de 169,96 observado neste estudo aproxima-se dos valores de referência (*baselines*) do estado da arte para a língua portuguesa reportados por [Marinho *et al.*, 2021] no corpus Essay-BR, onde modelos de regressão supervisionada obtiveram RMSE entre 159,44 e 163,92. Embora ligeiramente superior, o erro do ChatGPT é justificável pela ausência de treinamento específico (*fine-tuning*), demonstrando a capacidade de generalização da IAg sem a necessidade da complexa engenharia de atributos exigida pelos métodos clássicos.

5.1.3 Viés do modelo

Foi calculada a diferença média entre as notas atribuídas pelo ChatGPT e as atribuídas pelos avaliadores humanos. Valores positivos indicam que o modelo tende a atribuir notas maiores.

Tabela 3. Análise de viés: diferença média GPT - Humano

Competência	Viés (ChatGPT - Humano)
C1	12,74
C2	10,78
C3	25,99
C4	-0,47
C5	-10,98
Score Total	38,06

Análisando a Tabela 3, o modelo apresenta um viés de **superavaliação**, especialmente em C1, C2 e, principalmente, C3, onde o ChatGPT atribui em média 26 pontos a mais que os avaliadores humanos. Isso indica que o modelo valoriza excessivamente a construção argumentativa, mesmo quando ela não atende plenamente aos critérios do ENEM.

Por outro lado, há viés negativo em C4 e C5, sugerindo maior rigor na coesão e, principalmente, na proposta de intervenção. O viés positivo na nota total (+38 pontos) reforça o padrão de superavaliação geral.

5.1.4 Análise por faixa de desempenho

A análise por faixa revela que o alinhamento entre ChatGPT e avaliadores humanos varia conforme o nível da redação.

Faixa muito baixa

- MAE total: **210,7**
- Correlação: **0,60**
- Destaque positivo: modelo reconhece corretamente textos muito fracos (correlação moderada)
- Problema: erros elevados em todas as competências

Faixa baixa

- MAE total reduz significativamente (131,6), mas a correlação cai para **0,07**
- O modelo dificilmente distingue nuances dentro dessa faixa

Faixa média

- Menor MAE total: **111,5**
- Correlação da nota total: **inexistente (NaN)**
  Isso sugere dispersão insuficiente na amostra para cálculo de correlação ou baixa variabilidade.

Faixa alta

- MAE total: **82,5**
- Correlação: **0,26**
- Melhor desempenho relativo, mas ainda fraca concordância.

Faixa muito alta

- MAE volta a subir (**165,3**)
- Correlação melhora (**0,42**)
- Modelo tende a superavaliar textos já muito bons.

De forma geral, os resultados mostram que:

- O ChatGPT reproduz parcialmente o padrão geral de qualidade, mas não captura com fidelidade as especificidades de cada competência.
- Os maiores problemas estão na Competência 5 (Proposta de intervenção) — tanto em erro quanto em baixa correlação.
- O modelo apresenta viés consistente de superavaliação, sobretudo em C1, C2 e C3.
- A concordância por faixa é irregular, com destaque para: boa detecção de textos muito ruins (correlação 0,60); baixo alinhamento em quase todas as demais faixas; desempenho inconsistente em faixas média e alta.

Em conjunto, esses resultados indicam que o ChatGPT ainda não é confiável como avaliador automático direto, mas pode ser útil como ferramenta de apoio, especialmente para identificar erros gerais ou padrões argumentativos.

5.2 Análise dos Feedbacks

Além da avaliação numérica, esta pesquisa também investigou a qualidade dos feedbacks produzidos pelo ChatGPT. Para isso, utilizou-se prompt de validação responsável por analisar os textos gerados pelo modelo em quatro dimensões: **(i)** explicação do erro, **(ii)** identificação do que está faltando, **(iii)** sugestões de melhoria e **(iv)** aspectos positivos. Cada dimensão recebeu uma nota entre 0 e 5.

5.2.1 Estatísticas gerais dos feedbacks

A Tabela 4 apresenta as estatísticas descritivas das notas atribuídas pelo avaliador automático aos feedbacks.

Esses resultados indicam que o ChatGPT tende a produzir **feedbacks com bom nível de detalhamento**, especialmente na dimensão *Aspectos Positivos*, que apresentou a maior média (3,97). Isso sugere que o modelo adota uma **postura predominantemente encorajadora**, destacando qualidades do texto com maior frequência e intensidade.

Tabela 4. Estatísticas gerais dos feedbacks

Métrica	Média	Desvio padrão
Explicação do Erro	3,57	0,77
Apresentar o que está faltando	3,06	0,72
Pontos a melhorar	3,47	0,67
Aspectos positivos	3,97	0,74

A dimensão "Apresentar o que está faltando" apresentou a menor média (3,06), revelando que o modelo tem mais dificuldade em **identificar lacunas específicas** da redação, como elementos obrigatórios da tese, construção argumentativa ou estrutura da proposta de intervenção.

A Figura 2 mostra que a distribuição das notas é relativamente concentrada, com poucas notas inferiores a 2 e várias avaliações próximas de 4. Esse padrão evidencia uma **tendência geral do modelo a fornecer feedbacks positivos e de boa qualidade**, embora nem sempre suficientemente críticos ou direcionados.

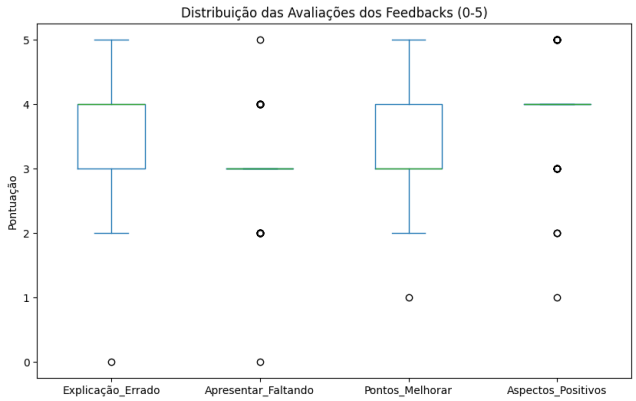

Figura 2. Distribuição das Avaliações dos Feedbacks

5.2.2 Qualidade dos feedbacks por faixa de desempenho da redação

A Tabela 5 apresenta a média das avaliações dos feedbacks segmentadas por faixa de desempenho da redação.

Tabela 5. Validação de Feedbacks por Faixa de Desempenho

Faixa	Explicação	Faltando	Melhorar	Positivos
muito baixa	3,30	2,70	3,05	3,40
baixa	3,60	3,10	3,30	3,80
média	3,75	3,00	3,60	4,05
alta	3,70	3,15	3,70	4,30
muito alta	3,50	3,35	3,70	4,30

Ao analisar essas faixas, observa-se que redações classificadas como *muito baixas* recebem os feedbacks menos completos, especialmente no critério *Apresentar o que está faltando*. Esse comportamento sugere que, diante de textos com muitos problemas estruturais, o modelo tende a produzir comentários mais gerais, com menor detalhamento sobre as lacunas específicas que deveriam ser preenchidas. Para redações classificadas como *médias* ou *altas*, os feedbacks tornam-se mais completos e consistentes. Nesses casos, o modelo apresenta maior capacidade de identificar erros e sugerir melhorias específicas, refletindo maior alinhamento com critérios avaliativos humanos. Um segundo padrão observado é a tendência do ChatGPT a enfatizar aspectos positivos em todas as faixas, com as maiores médias nas faixas *alta* e *muito alta*. Esse comportamento reforça a postura predominantemente motivadora do modelo, que prioriza destacar qualidades mesmo quando as demais dimensões ainda apontam limitações.

5.2.3 Análise qualitativa a partir de sentimentos

Para complementar a interpretação dos feedbacks **gerados pelo ChatGPT**, realizou-se uma análise de sentimento utilizando o modelo *BERT multilingual*[1], capaz de classificar textos em uma escala de 1 a 5 estrelas, representando desde avaliações muito negativas até avaliações muito positivas. Esse procedimento permite identificar a percepção geral dos validadores sobre a qualidade dos feedbacks gerados pela ferramenta.

A classificação foi automatizada via script em linguagem Python, utilizando a biblioteca Transformers. O processo consistiu em submeter os textos das análises qualitativas ao modelo, que calculou a probabilidade de cada comentário pertencer a uma das cinco categorias de sentimento. Ao final, atribuiu-se a cada feedback a nota (estrela) correspondente à categoria de maior probabilidade identificada pelo sistema.

Os resultados indicam que a maior parte das análises qualitativas apresenta caráter neutro (51,58%), sugerindo que os avaliadores tendem a adotar uma postura descritiva e técnica ao comentar sobre o feedback, sem expressar julgamentos emocionais fortes. Além disso, observa-se uma proporção significativa de avaliações positivas (47,36%), nas quais os validadores reconhecem elementos construtivos, clareza textual ou pertinência das observações fornecidas pelo sistema.

A presença de avaliações negativas é mínima (1,05%), e não foram identificadas avaliações extremamente negativas ou extremamente positivas (1 ou 5 estrelas). Esse comportamento indica que, embora existam críticas pontuais — geralmente associadas à superficialidade ou à falta de detalhamento em alguns casos — a percepção geral dos avaliadores é favorável ao conteúdo gerado.

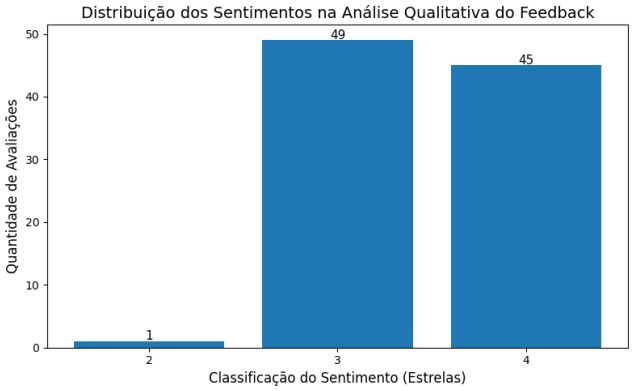

Figura 3. Distribuição da classificação de sentimento atribuída às análises qualitativas dos feedbacks gerados pelo modelo. A maior parte das avaliações encontra-se entre 3 e 4 estrelas, indicando percepção predominantemente neutra ou positiva por parte dos validadores.

[1]Modelo disponível em: `https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment`

De forma geral, como mostrado na Figura 3 a análise de sentimento revela que os feedbacks produzidos pelo modelo são predominantemente neutros ou positivos, sendo raramente avaliados como insatisfatórios. Esse resultado reforça a conclusão de que o sistema é capaz de fornecer orientações úteis e pedagogicamente válidas, ainda que haja espaço para aprimoramentos em profundidade e especificidade.

5.3 Divergência dos Resultados

Esta seção analisa a variabilidade dos resultados produzidos pelo modelo **ChatGPT 3.5-turbo** na avaliação automática de redações, considerando múltiplas execuções independentes com **dados de entrada idênticos** e **prompt inalterado**. O objetivo é investigar o grau de consistência das notas e identificar possíveis padrões de divergência inerentes ao comportamento estocástico do modelo. Foram realizadas **três execuções independentes** *(runs)* do processo de avaliação, mantendo-se constantes: (i) o conjunto de redações analisadas, (ii) o prompt de avaliação baseado nos critérios do ENEM, e (iii) o modelo utilizado. As respostas do modelo foram processadas automaticamente, estruturadas em *DataFrames* e exportadas em formato CSV, permitindo a comparação direta entre as execuções. Optou-se por utilizar a temperatura padrão (default) para simular o cenário real de uso de um estudante ou professor leigo, que utilizaria a interface web ou configurações padrão, sujeitando-se à estocasticidade natural da ferramenta.

5.3.1 Notas avaliativas

A análise quantitativa concentrou-se na **nota total** e nas **cinco competências avaliadas**. Para cada critério, foram calculadas as seguintes métricas: **desvio padrão médio**, **amplitude média** e **concordância exata**, definida como a porcentagem de casos em que as três execuções produziram exatamente o mesmo valor. Os resultados indicam **variação relevante** entre as execuções. A nota total apresentou o maior nível de divergência, com **desvio padrão médio superior a 100 pontos** e **amplitude média próxima a 200 pontos**, enquanto as competências individuais exibiram desvios menores, porém ainda relevantes. A **concordância exata** foi extremamente baixa em todos os critérios, variando entre **0% e 2%**, o que evidencia a ausência de determinismo nas respostas do modelo, como detalha a Tabela 6.

Tabela 6. Divergência de Notas

Critério	STD	Concordância Exata (%)
score_GPT	103.23	0.0%
C1_GPT	20.40	1.0%
C2_GPT	23.18	1.0%
C3_GPT	25.58	0.0%
C4_GPT	26.72	2.0%
C5_GPT	33.91	1.0%

5.3.2 Avaliação dos feedbacks - Feedbacks distintos

Para analisar a consistência na avaliação da qualidade dos feedbacks gerados, seguiu-se o procedimento da etapa anterior. Foram realizadas três execuções independentes do processo de validação, mantendo-se constantes as redações, o prompt e o modelo utilizado. Ressalta-se que os feedbacks avaliados foram gerados de forma independente a cada execução, ou seja, não se tratam de avaliações repetidas de um mesmo feedback. O fato de gerar um feedback a cada interação é importante pois queremos validar a consistência de um bom feedback e não a capacidade do ChatGPT avaliar o feedback em si.

As avaliações consideraram uma escala discreta de 0 (zero) a 5 (cinco) para cada critério, o que reduz a margem de oscilação numérica e favorece maior estabilidade nas comparações. Os resultados indicam **menor variabilidade** quando comparados à etapa de atribuição de notas às redações. Os desvios padrão médios situaram-se entre 0,53 e 0,77 pontos, evidenciando variações moderadas dentro da escala adotada.

Como mostra a Tabela 7, a **concordância exata** ficou entre **15% e 29%** das avaliações, o que sugere maior estabilidade do modelo ao julgar a qualidade pedagógica dos feedbacks do que ao atribuir notas numéricas globais. Destaca-se que essa divergência reflete a avaliação de feedbacks textualmente distintos, porém semanticamente equivalentes, gerados em execuções independentes, indicando consistência nos critérios avaliativos, mesmo diante de variações na geração textual.

Tabela 7. Divergência de Qualidade dos Feedbacks

Métrica	STD	Concordância Exata (%)
Explicação do Erro	0.7668	27.9%
Apresentar o que está faltando	0.5390	29.1%
Pontos a melhorar	0.6094	15.1%
Aspectos positivos	0.6352	16.3%

6 Conclusão

O presente trabalho investigou a viabilidade do uso de LLMs, especificamente o ChatGPT, como ferramenta de apoio à correção e geração de feedbacks para redações de alunos do ensino médio, seguindo os critérios do ENEM. A motivação central residiu na necessidade de oferecer ganho de escala ao processo avaliativo, mitigando a sobrecarga docente e proporcionando retornos mais ágeis aos estudantes.

Os resultados obtidos através da análise de 100 redações do corpus *Essay-BR* demonstraram um cenário divergente. Do ponto de vista **quantitativo**, a correlação moderada (r $\approx 0,45\%$) entre as notas atribuídas pela IA e as notas dos avaliadores humanos, somada a um MAE de aproximadamente 140 pontos na nota total, indicam que o modelo — em sua configuração atual de *prompting* (RACE) — ainda não possui precisão suficiente para substituir a avaliação humana. Observou-se, inclusive, um viés positivo na avaliação da IA, que tendeu a atribuir notas superiores às dos avaliadores oficiais, com destaque para a dificuldade do modelo em penalizar corretamente desvios na Competência 1 (Norma Culta) e na Competência 5 (Proposta de Intervenção).

Entretanto, sob a ótica **qualitativa e formativa**, a ferramenta demonstrou alto potencial. A validação via autoavaliação indicou que os feedbacks gerados são estruturalmente coerentes e polidos. A análise de sentimento (via BERT) e as métricas de qualidade do feedback (média de 3,97/5 para

aspectos positivos) sugerem que a IA é capaz de produzir comentários encorajadores e explicativos. Embora a pontuação numérica precise de refinamento, a capacidade do modelo de identificar o tema e estruturar uma resposta pedagógica é promissora para o uso como "tutor inteligente" em ambientes de aprendizado, servindo como uma primeira camada de revisão para o aluno antes da submissão ao professor.

Conclui-se, portanto, que a IA Generativa, no estágio tecnológico avaliado, atua melhor como um **instrumento de suporte educacional** do que como um **avaliador independente**. Ela pode democratizar o acesso a feedbacks detalhados, desde que supervisionada.

Limitações da Pesquisa: Deve-se ressaltar que a utilização da própria ferramenta para validar a qualidade de seus feedbacks (autoavaliação) constitui uma limitação deste estudo. Embora adotada em caráter exploratório, essa abordagem pode incorrer na reprodução de vieses inerentes ao modelo, não substituindo a necessidade de validação por especialistas humanos para resultados definitivos.

Como trabalhos futuros, sugerem-se:

1. Utilização de diferentes ferramentas de IAg, principalmente, especificamente treinadas para ambientes educacionais.
2. Inferir a capacidade de feedback em diferentes níveis de ensino para diferentes conteúdos.
3. Buscar refinar os prompts em nível de detalhe.
4. Validação dos feedbacks gerados através de avaliação humana realizada por docentes da área de linguagens.

Declarações complementares

Contribuições dos autores

Luiz Oliveira é o principal contribuidor e escritor deste manuscrito. Taciana Pontual Falcão é a orientadora do trabalho, responsável pela revisão do texto e orientações de metodologia de pesquisa.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Disponibilidade de dados e materiais

Os conjuntos de dados (e/ou softwares) gerados e/ou analisados durante o estudo atual estão disponíveis no `Google Colab`[2] e no `Google Drive`[3].

Outras informações relevantes

Uso de IA Generativas como ChatGPT e Claude para validação gramatical e suporte na codificação para análise dos resultados.

Referências

Adelina Moura, A. A. C. (2023). Literacia de prompts para potenciar o uso da inteligência artificial na educação. *RE@ D-Revista de Educação a Distância e Elearning*, 6(2):e202308–e202308. DOI: 10.34627/redvol6iss2e202308.

Costa Júnior, J. F., Moraes, L. S., de Souza, M. M. N., Lopes, L. C. L., Meneses, A. R., Pontes Pinto, A. R. d. A., dos Santos, L. S. R., and Zocolotto, A. (2023). A importância de um ambiente de aprendizagem positivo e eficaz para os alunos. *Rebena - Revista Brasileira de Ensino e Aprendizagem*, 6:324–341.

Cunto, C., Fernandes, A., Pinto, L., Ferreira, J., Souza, W., Rodrigues, S., Silva, W., and Angonese, A. (2022). O feedback personalizado por meio de software de avaliaÇÃo de atividades qualitativas. *Anais do CIET:CIESUD:2022*.

da Silva Neta, M. and Capuchinho, A. C. (2017). Educação híbrida: Conceitos, reflexões e possibilidades do ensino personalizado. In *II Congresso sobre Tecnologias na Educação (Ctrl+E 2017)*, pages 148–156. CEUR Workshop Proceedings.

Di Santo, M. (2023). Inteligência artificial na educação: Rumo a uma aprendizagem personalizada. *IOSR Journal of Humanities and Social Science*, 28:19–25. DOI: 10.9790/0837-2805031925.

Giraffa, L. and Kohls-Santos, P. (2023). Inteligência artificial e educação: conceitos, aplicações e implicações no fazer docente. *Educação em Análise*, 8(1):116–134. DOI: 10.5433/1984-7939.2023v8n1p116.

Guerreiro, N. P., Nunes, E. d. F. P. d. A., González, A. D., and Mesas, A. E. (2016). Perfil sociodemogrÁfico, condiÇÕes e cargas de trabalho de professores da rede estadual de ensino de um municÍpio da regiÃo sul do brasil. *Trabalho, Educação e Saúde*, 14:197–217. DOI: 10.1590/1981-7746-sol00027.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-eval: Nlg evaluation using gpt-4 with better human alignment. DOI: 10.48550/arXiv.2303.16634.

Marinho, J., Anchiêta, R., and Moura, R. (2021). Essay-br: a brazilian corpus of essays. In *Anais do III Dataset Showcase Workshop*, pages 53–64, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/dsw.2021.17414.

Picão, F., Gomes, L., Alves, L., Barpi, O., and Luccheti, T. (2023). InteligÊncia artificial e educaÇÃo: Como a ia estÁ mudando a maneira como aprendemos e ensinamos. *Revista Amor Mundi*, 4:197–201. DOI: 10.46550/amor-mundi.v4i5.254.

Pontual Falcão, T., Arêdes, V., de Souza, S. B. J., Fiorentino, G., Neto, J. R., Alves, G., and Mello, R. F. (2023). Tutoria: a software platform to improve feedback in education. *Journal on Interactive Systems*, 14(1):383–393. DOI: 10.5753/jis.2023.3247.

Silva, A. S. d., Alarcão, D. T. A., and Faria, S. P. (2025). A inteligÊncia artificial como facilitadora no ensino personalizado: Potencialidades e desafios no contexto educacional brasileiro. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, 11(6):3474–3499. DOI: 10.51891/rease.v11i6.19991.

Zhu, Y., Zhang, L., Zong, R., and Sun, H. (2024). Exploring ai-generated feedback on english writing: A case study of chatgpt. *US-China Foreign Language*, 22(3):144–153. DOI: 10.17265/1539-8080/2024.03.002.

[2]URL: https://colab.research.google.com/drive/1l_wehdGI6_knaBJsJu_dS4AAG3NLZYGU

[3]URL: https://drive.google.com/drive/folders/1TJLcfzMROpE9mVnUgRv2jRyl3rS2SzVm