

## **Ensemble de LLMs: Aliando a eficiência da IA com a expertise humana na avaliação de redações**

Hugo Kenji Pereira Harada<sup>1</sup>, Marcos Alexandre Magri<sup>2</sup>, Heliton Ribeiro Tavares<sup>3</sup>

<sup>1</sup> Adaptativa Inteligência Educacional Ltda. hugo.harada@adaptativa.com.br

<sup>2</sup> Adaptativa Inteligência Educacional Ltda. marcos.magri@adaptativa.com.br

<sup>3</sup> Universidade Federal do Pará. heliton@ufpa.br

### **Resumo**

Este estudo explora o potencial dos Grandes Modelos de Linguagem (LLMs) na correção de redações, propondo uma abordagem ensemble para mitigar inconsistências e gerar avaliações mais confiáveis. Sete redações de estudantes do ensino médio foram avaliadas por LLMs das empresas OpenAI, Anthropic e Google, utilizando uma rubrica detalhada. As correções foram submetidas a três algoritmos de arbitragem para consolidação das notas. Os resultados indicam que, apesar de variações, os LLMs mantiveram um nível aceitável de consistência com a avaliação humana, não excedendo o limite de qualidade de 80 pontos estabelecido pelo INEP/MEC. A abordagem ensemble pode ter contribuído para esse resultado. Conclui-se que os LLMs têm potencial para auxiliar na avaliação de redações, respeitando parâmetros estabelecidos. Uma abordagem híbrida, combinando avaliação preliminar dos LLMs com refinamento por professores, é sugerida para potencializar a eficiência do processo de correção, preservando o julgamento humano em aspectos mais subjetivos.

**Palavras-chave:** Chatgpt, Grandes Modelos de Linguagens, LLMs, Redação, Ensemble

### **1. Introdução**

Os Grandes Modelos de Linguagem (Large Language Models, LLMs) apresentam um potencial significativo para impactar positivamente a vida acadêmica dos estudantes quando utilizados na correção de redações. Um dos principais benefícios é a capacidade dos LLMs de fornecer feedback detalhado, específico e individualizado para cada aluno quase em tempo real, destacando os pontos fortes e identificando as áreas que necessitam de aprimoramento em suas produções textuais [1]. Isso permite que os discentes recebam orientações valiosas sobre como melhorar suas habilidades de escrita, mesmo em contextos nos quais os docentes não dispõem de tempo para oferecer feedback individualizado e detalhado a cada estudante. Adicionalmente, ao automatizar parte do processo de correção de redações, os LLMs podem liberar tempo valioso dos professores. Consequentemente, os educadores podem concentrar-se em atividades de alto valor agregado, como elaborar estratégias pedagógicas engajadoras, ministrar aulas

interativas e fornecer suporte individualizado aos alunos que mais necessitam. Esse uso eficiente do tempo docente tem o potencial de aprimorar a qualidade geral do ensino e da aprendizagem. Estudos também evidenciam que o feedback gerado por LLMs em redações pode aumentar a motivação e promover emoções positivas nos alunos em relação à escrita [2]. Esse engajamento intensificado com a escrita, impulsionado pelo feedback dos LLMs, tem o potencial de acelerar significativamente o desenvolvimento das competências de comunicação escrita dos discentes ao longo de sua trajetória acadêmica.

No entanto, apesar de todo esse potencial, o desempenho dos LLMs como corretores de redação pode variar consideravelmente dependendo de fatores como o tamanho da janela de contexto, a escala dos dados de treinamento e a qualidade dos prompts utilizados [3]. Isso pode levar a correções inconsistentes, mesmo quando se utiliza um prompt bem elaborado baseado em rubricas estabelecidas.

Para mitigar esse problema, propomos uma abordagem ensemble em que a redação de um aluno é submetida a múltiplos LLMs independentes, cada um gerando sua própria correção e nota. Essas avaliações individuais são então consolidadas por um módulo "árbitro" que determina a nota final da redação conforme a Figura 1. Serão testados três algoritmos diferentes para esse módulo de arbitragem:

1. Seleção da Melhor Correção: O árbitro analisa as correções submetidas e escolhe aquela com melhor adequação às rubricas de correção.
2. Derivação de Nova Correção: O árbitro sintetiza uma nova correção a partir das submissões individuais, buscando incorporar os pontos fortes de cada uma.
3. Calibração com Redações Previamente Avaliadas: Semelhante ao algoritmo 2, mas a correção final gerada pelo árbitro é então calibrada com base em um conjunto de redações previamente avaliadas por professores humanos para garantir consistência com os padrões estabelecidos.

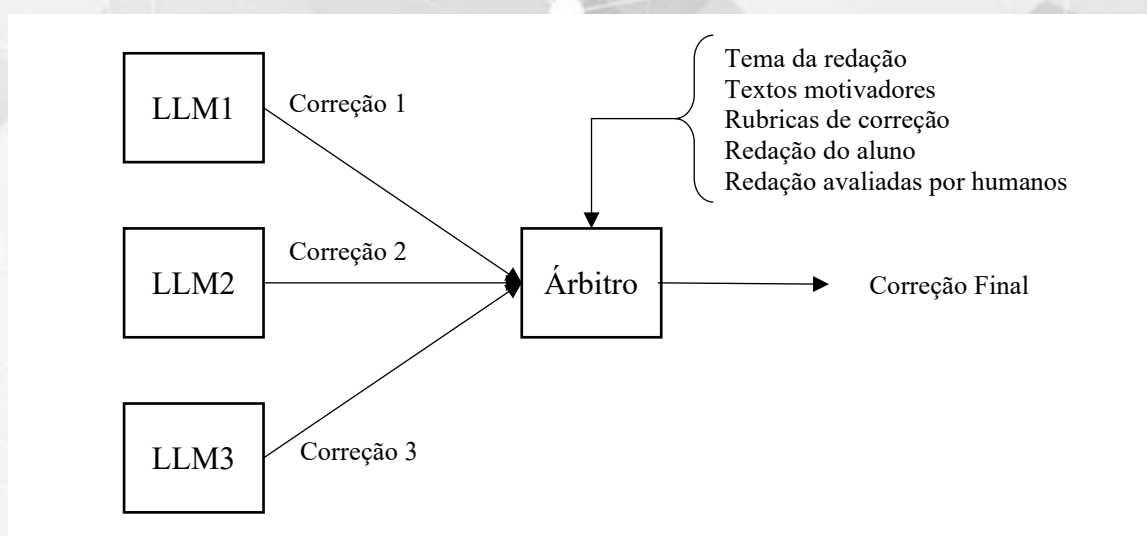


Figura 1 - Correção de redações por abordagem Ensemble

Acredita-se que essa abordagem ensemble, ao aproveitar as capacidades de múltiplos LLMs e mecanismos de consolidação robustos, pode gerar correções mais confiáveis e consistentes em comparação com a dependência de um único modelo.

## 2. Objetivos

O principal objetivo deste estudo é comparar o desempenho de corretores de redações baseados em um único modelo de linguagem (LLM) com uma abordagem de ensemble, na qual múltiplos LLMs são utilizados para gerar correções mais consistentes e precisas. Pretende-se avaliar se a abordagem ensemble, ao combinar diferentes correções, pode superar as limitações de inconsistência observadas em corretores automatizados baseados em um único modelo. Além disso, será verificada a coerência entre as notas atribuídas pelas LLMs e pelos corretores humanos, bem como a eficácia dos três algoritmos de arbitragem propostos para consolidar as correções.

## 3. Metodologia

Sete redações de alunos do 3º ano do ensino médio foram selecionadas, apresentando diferentes níveis de qualidade, todas sobre tema "O estilo de vida moderno e seus efeitos sobre a saúde humana". Essas redações foram previamente avaliadas por corretores humanos e foram avaliadas por modelos de linguagem de grande porte (LLMs) disponibilizados pelas empresas OpenAI [2], Anthropic [3] e Google [4] conforme a Tabela 1. O prompt utilizado para a correção pelas LLMs implementou a rubrica detalhada na Tabela 2 [7] e incluiu também o tema da redação, os textos motivadores e orientações específicas sobre a formatação da nota e do feedback.

As correções realizadas pelos LLMs foram submetidas a três árbitros, que geraram novas avaliações seguindo os três algoritmos detalhados anteriormente. Cada algoritmo foi aplicado utilizando os três LLMs disponíveis, totalizando 12 correções adicionais para cada redação, além da avaliação humana inicial. A correção por inteligência artificial foi considerada coerente se a nota atribuída não diferisse em mais de 80 pontos da nota dada pelos corretores humanos. Esse é o mesmo critério utilizado pelo INEP/MEC para acionar uma terceira correção caso haja discrepância maior que 80 pontos na avaliação de dois corretores independentes para uma mesma competência [6].

Tabela 1 Especificações dos LLMs

Empresa	Modelo	Tamanho da janela de Contexto	Número máximo de tokens de saída
OpenAI	gpt-4o-2024-05-13	128.000 tokens	4.000 tokens
Anthropic	claude-3-5-sonnet-20240620	200.000 tokens	8.000 tokens
Google	gemini-1.5-pro	2.097.152 tokens	8192 tokens

Tabela 2 - Rubrica de correção da competência 3 de coerência da redação do ENEM

CRITÉRIO	200 PONTOS	160 PONTOS	120 PONTOS	80 PONTOS	40 PONTOS	0 PONTOS
As informações são apresentadas de modo a defender adequadamente um ponto de vista claro ao longo do texto, ou seja, elas não possuem contradições entre si e foram ordenadas de forma coerente?	Sim, pois as informações selecionadas não são contraditórias e foram ordenadas de forma coerente, de modo a defender um ponto de vista claro ao longo do texto, que não se limita ao senso comum sobre o tema.	Sim, pois as informações selecionadas não são contraditórias e foram ordenadas de forma coerente, de modo a defender um ponto de vista claro ao longo do texto, mas esse se limita ao senso comum sobre o tema.	Sim, pois as informações selecionadas não são contraditórias, mas foram ordenadas de modo confuso, o que prejudica a qualidade da argumentação em defesa do ponto de vista central do texto.	Não, pois as informações apresentam contradições e/ou foram ordenadas de modo confuso, o que prejudica a qualidade da argumentação em defesa do ponto de vista central do texto.	Não, pois o autor não defende um ponto de vista claro ao longo do texto.	Não, pois as informações são apresentadas de forma desconexa, não configurando um texto.
As informações selecionadas são coerentes com o tema da proposta e demonstram o bom repertório cultural do autor?	Sim, pois as informações estão de acordo com o tema, não foram sugeridas pelos textos motivadores e não pertencem ao senso comum.	Sim, pois as informações estão de acordo com o tema, mas grande parte delas pertence ao senso comum.	Sim, pois as informações estão de acordo com o tema, mas a maior parte delas pertencem ao senso comum.	Sim, pois as informações estão de acordo com o tema, mas consistem em observações genéricas e típicas do senso comum.	Não, pois o assunto geral é abordado, mas ocorre tangenciamento de tema.	Não, pois o autor não aborda o assunto exigido e ocorre a fuga do tema.

#### 4. Resultados e Discussão

A análise da Tabela 3, que compara as notas atribuídas por corretores humanos e por diferentes modelos de linguagem (LLMs) para a competência 3 (coerência) em redações de estudantes, revela um achado importante: em nenhum caso as notas dos LLMs excederam o limite de qualidade de 80 pontos em relação à avaliação humana, critério estabelecido pelo INEP/MEC para acionar uma terceira correção em caso de discrepância.

Esse resultado é encorajador, pois indica que, apesar das variações observadas, os modelos de linguagem conseguiram manter um nível aceitável de consistência com o julgamento dos corretores humanos. O fato de todos os LLMs respeitarem o limite de 80 pontos sugere que eles são capazes de capturar, até certo ponto, os critérios essenciais utilizados pelos humanos na avaliação da coerência textual.

No entanto, mesmo dentro desse limite, ainda foram observadas diferenças consideráveis entre as notas atribuídas pelos diferentes LLMs e entre elas e a avaliação humana. Apesar da proximidade nas notas, os LLMs tendem a não capturar com precisão os extremos de qualidade textual. Em redações de maior qualidade (como a ID 233534), há uma tendência de os modelos subavaliarem certos aspectos, especialmente no que diz respeito a nuances mais sofisticadas da argumentação. Isso reflete uma limitação dos LLMs em alcançar o mesmo nível de percepção detalhada que um corretor humano treinado pode identificar. Para redações com notas mais baixas (como ID 233072), os LLMs também apresentaram variações controladas, mantendo-se dentro da faixa de 80 pontos. Isso sugere que, mesmo diante de textos com falhas claras de coerência, os modelos conseguiram identificar essas inconsistências, embora algumas pequenas variações entre os modelos possam ter surgido devido a diferenças na forma como cada LLM processa a falta de clareza argumentativa.

É possível que a abordagem de ensemble, que combina múltiplos LLMs por meio de técnicas de arbitragem, tenha contribuído para manter as notas dentro do limite aceitável. Ao compensar eventuais desvios de modelos individuais, o ensemble pode ter favorecido uma maior consistência com a avaliação humana.

Para validar a generalização desses achados, seria interessante expandir a análise para uma amostra maior de redações e verificar se o padrão de respeito ao limite de qualidade se mantém. Além disso, investigar se esse comportamento também é observado na avaliação das demais competências poderia fornecer uma visão mais abrangente do desempenho dos LLMs.

## 5. Conclusões

O cumprimento consistente do limite de qualidade pelos LLMs é um resultado promissor que merece destaque na discussão dos resultados. Ele indica o potencial dessas tecnologias para auxiliar na avaliação de redações de forma confiável, respeitando os parâmetros estabelecidos pelo INEP/MEC. Esse achado fortalece a perspectiva de que, com aprimoramentos contínuos, os modelos de linguagem podem se tornar ferramentas valiosas para apoiar o processo de avaliação, oferecendo aos corretores humanos uma segunda opinião consistente e embasada em critérios objetivos.

Além disso, em vez de substituir os professores no processo de correção, a saída gerada pelos LLMs poderia funcionar como o primeiro passo em uma avaliação humana. Os modelos podem fornecer uma avaliação preliminar objetiva e consistente, servindo como uma base sobre a qual os professores podem construir e refinar sua análise, garantindo que aspectos mais subjetivos e nuances de alta complexidade sejam abordados por corretores humanos. Essa abordagem híbrida potencializaria a eficiência do processo de correção, ao mesmo tempo em que preserva o julgamento humano, essencial em questões mais delicadas e interpretativas da avaliação de redações.

Tabela 3 Resultado da avaliação da competência 3 para redações de estudantes

ID	Nota Redação (Humano)		Nota Competência 3 Humano	OpenAI	Anthropic	Google
233534	1000	Correção	200	160	200	120
		Melhor	-	Anthropic	Anthropic	Anthropic
		Derivada	-	160	160	160
		Calibrada	-	160	160	160
233967	960	Correção	200	160	160	120
		Melhor	-	Anthropic	Anthropic	Anthropic
		Derivada	-	160	160	160
		Calibrada	-	160	160	160
233200	800	Correção	160	160	200	120
		Melhor	-	Anthropic	Anthropic	Anthropic
		Derivada	-	160	160	160
		Calibrada	-	160	160	160
233600	680	Correção	120	160	160	120
		Melhor	-	OpenAI	Anthropic	Anthropic
		Derivada	-	160	160	160
		Calibrada	-	160	160	160
260235	520	Correção	120	160	120	120
		Melhor	-	Anthropic	Anthropic	Anthropic
		Derivada	-	120	120	120
		Calibrada	-	120	120	160
235917	440	Correção	80	160	160	80
		Melhor	-	Anthropic	Anthropic	Anthropic
		Derivada	-	160	160	120
		Calibrada	-	160	120	120

233072	360	Correção	80	80	120	120
		Melhor	-	Anthropic	Anthropic	Anthropic
		Derivada	-	120	120	120
		Calibrada	-	120	120	80

## 6. Referências

- [1] TANG, X., et al. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon* 10. Cell Press.
- [2] MEYER et al. (2024). Using LLMs to bring evidence-based feedback into the classroom. *Computers and Education: Artificial Intelligence*.
- [3] STAHL, Maja; BIERMANN, Leon; NEHRING, Andreas; WACHSMUTH, Henning. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. In: *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, June 20, 2024. Association for Computational Linguistics, 2024, p. 283-298.
- [4] OpenAI. (2023). ChatGPT [Modelo de linguagem AI]. Disponível em <https://chat.openai.com>. Acesso em: 10 set 2024.
- [5] Anthropic. (2023). Claude API: A conversational AI assistant [Computer software]. Disponível em: <https://www.anthropic.com>. Acesso em: 10 set 2024.
- [6] GOOGLE AI. Gemini. [S.l.], [s.d.]. Disponível em: <https://gemini.google.com>. Acesso em: 10 set 2024.
- [7] MAGRI, M. A Redação no ENEM. - São Paulo, SP: Ensinar, 2015. 1. edição. 2. edição, 2023. 3. edição, 2024. ISBN 978-85-60985-61-6
- [8] INEP/MEC. A redação do enem 2023 - Cartilha do participante.