Proposta recebida em Abril 2026 e aceite para publicação em Setembro 2026.

Avaliação de Questões Discursivas Curtas por LLMs: Desafios e Potenciais Aplicações Educacionais

Assessment of Short Essay Questions by LLMs: Challenges and Potential Educational Applications

Roberlei Alves Bertucci
Universidade Tecnológica Federal do Paraná

Osmar de Oliveira Braz Junior
Universidade do Estado de Santa Catarina

Renato Fileto
Universidade Federal de Santa Catarina

Resumo

A relação entre tecnologias emergentes e linguagem humana tem sido amplamente investigada, especialmente após a popularização de grandes modelos de linguagem (LLMs, do inglês *Large Language Models*) e de sua capacidade de interpretar e gerar textos. No contexto educacional, há intenso debate sobre o uso dessas tecnologias em sala de aula, seja por estudantes, docentes ou instituições. Considerando os avanços no uso de inteligência artificial (IA) na linguística computacional, a pesquisa justifica-se pela necessidade de integrar metodologias inovadoras na análise textual, mais especificamente, na avaliação de questões discursivas curtas (ASAG, do inglês, *Automatic Short Answer Grading*), respondidas por estudantes de graduação em uma atividade de interpretação textual. O estudo compara a avaliação humana, marcada por subjetividade, com as capacidades dos LLMs, cujo funcionamento se baseia em cálculos probabilísticos. A análise dos dados revelou um padrão de notas bastante próximo entre os dois avaliadores humanos, enquanto os LLMs apresentaram maior variação, tanto entre si quanto em relação aos humanos. A maior correlação positiva entre um LLM e os avaliadores humanos foi observada com o Gemini 3, em uma das variações de *prompt* propostas neste estudo. Esses achados indicam que ainda são necessários mais estudos para refinar a tarefa de ASAG em português. Ainda assim, desempenhos como os observados com o Gemini sugerem potencial de contribuição educacional, especialmente no apoio ao diagnóstico e ao *feedback* docente em questões discursivas curtas.

Palavras chave

correção automática de respostas discursivas curtas, ASAG, avaliação das respostas dos alunos, grandes modelos de linguagem, LLM

Abstract

The relationship between emerging technologies and human language has been widely investigated, especially after the popularization of large language models (LLMs) and their ability to interpret and generate texts. In educational settings, there is intense debate about the use of these technologies in classrooms by students, teachers, and institutions. Considering recent advances in the use of Artificial Intelligence (AI) in computational linguistics, this research is motivated by the need to integrate innovative methodologies into textual analysis, specifically in the assessment of short-answer questions (ASAG, Automatic Short Answer Grading) answered by undergraduate students in a text interpretation activity. This study compares human evaluation, marked by subjectivity, with the capabilities of LLMs, whose operation is based on probabilistic calculations. Data analysis revealed a very similar grading pattern between the two human evaluators, whereas the LLMs showed greater variation, both across models and in comparison with human evaluators. The highest positive correlation between an LLM and human evaluators was observed for Gemini 3, in one of the prompt variations tested in this study. These findings indicate that further research is still needed to refine tools for the ASAG task in Portuguese. Nevertheless, the observed performance of models such as Gemini suggests potential educational contributions, particularly in supporting teachers' diagnosis and feedback practices in short-answer assessment.

Keywords

automatic short-answer grading, ASAG, evaluation of student answers, large language models, LLM

1. Introdução

A educação brasileira tem enfrentado grandes desafios ao longo dos anos, sobretudo no que diz respeito à interpretação e produção de textos, conforme se vê retratado nos dados de avaliações oficiais (Peduzzi, 2024; Presidência da República, 2025; TPE, 2024). Pesquisas recentes também

mostram que os professores em formação têm dificuldades em tarefas de leitura e interpretação de texto, ainda que o auxílio de estratégias — como o uso de anotações durante a leitura — possa contribuir para amenizar o problema (Ferreira & Cafardo, 2026; Nunes & Bertucci, 2025; UTFPR, 2025).

Nesse cenário, destacamos também a importância de se avaliar a leitura na escola, uma vez que isso pode ser um instrumento de diagnóstico do desempenho dos estudantes e, em alguma medida, orientar as políticas de ensino no tema (Coscarelli & Prazeres, 2013). A avaliação do desempenho de leitura ocorre por meio de tarefas que se propõem a medir (ainda que subjetivamente) o quanto um determinado estudante compreendeu de um determinado texto. Quando objetivas, essas tarefas podem ser tratadas computacionalmente, inclusive com métricas que valorizam o raciocínio e punem a resposta aleatória (Teoria de Resposta ao Item Lalor et al. (2024), por exemplo). Já as questões discursivas exigem a integração de diferentes aspectos para a checagem de sua qualidade e questões como a subjetividade humana (Cançado et al., 2020) ou vieses de interpretação de técnicas computacionais (Ye et al., 2025) podem interferir no grau de avaliação do texto, o que precisa ser amenizado, sobretudo em processos de diagnóstico e avaliação.

Diferentes trabalhos analisam a contribuição de ferramentas computacionais para a avaliação de textos como forma de apoiar o ensino (Galhardi et al., 2022; Rassi & Lopes, 2023; Monteiro & Barreto, 2022). Nesse contexto, o uso de soluções automatizadas tem potencial para produzir efeitos práticos capazes de auxiliar tanto o docente quanto o discente: é o que nos mostram, por exemplo, Monteiro & Barreto (2022), que implementam *feedback* baseado na Teoria de Resposta ao Item para as questões objetivas do Exame Nacional do Ensino Médio (ENEM), favorecendo o estudo dos discentes e o diagnóstico dos docentes. Taschetto & Fileto (2025) mostram que grandes modelos de linguagem (LLMs, do inglês *Large Language Models*) já atingem desempenho superior à grande maioria dos humanos no ENEM, tanto em termos de medidas de desempenho tradicionalmente usadas para avaliar sistemas, tais como a acurácia, quanto quando avaliados usando a Teoria de Resposta ao Item, principalmente quando utilizadas técnicas como geração aumentada por recuperação (Fan et al., 2024). Braz Junior et al. (2025) mostram que, embora não superem a avaliação humana, LLMs podem auxiliar na avaliação de resumos, por meio do emprego de verbos *dicendi*.

Além disso, questões discursivas e redações também têm sido objeto de investigação para correção automática, como se observa em trabalhos em português (Galhardi et al., 2018, 2020, 2022; Meira et al., 2023; Rassi & Lopes, 2023; Silva, 2023; Almeida & Moura, 2024; Santos, 2025; Oliveira Júnior, 2025). Especificamente, o uso de LLMs tem sido visto como uma alternativa a estratégias tradicionais, porque, como afirmam Ferreira Mello et al. (2025), eles são treinados com um conjunto enorme de dados e têm demonstrado bom desempenho em tarefas de resposta a diferentes temas, o que os torna candidatos a aplicações educacionais, como é o caso da avaliação de questões discursivas curtas (ASAG, do inglês, *Automatic Short Answer Grading*).

Nesse contexto, esta pesquisa investiga a seguinte questão: quão adequada é a avaliação realizada por LLMs para questões discursivas curtas em comparação com a avaliação humana? Para responder a essa questão, tomamos um conjunto de 129 respostas dadas por estudantes de graduação, as quais foram corrigidas por dois avaliadores humanos. A seguir, submetemos esses mesmos 129 textos à avaliação automática dos LLMs, com algumas variações nos *prompts*, conforme será detalhado na Seção 4. Com base nisso, o artigo apresenta as seguintes contribuições:

1. *Corpus* diferente: ao focar na avaliação de uma questão curta e objetiva, a pesquisa foge das tradicionais análises sobre redações, encontradas na área de correção automática em português; além disso, apresenta dados de estudantes de graduação (professores em formação), um grupo que também é pouco enfocado em trabalhos sobre leitura e escrita.
2. Abordagem comparativa: ao comparar a avaliação humana com a de LLMs, propomos um olhar integrado do papel que essas tecnologias podem oferecer ao ensino.
3. Variação de *prompts* e modelos: justamente por trabalhar com LLMs, a pesquisa avalia o impacto que mudanças no *prompt* podem acarretar na tarefa de ASAG, tal como se indica em pesquisas com modelos de linguagem; além disso, oferece resultados sobre como cinco modelos diferentes se comportam diante de uma mesma tarefa.

Assim, nesta pesquisa, considera-se que a avaliação automática de textos pode auxiliar tanto o diagnóstico do professor (e a consequente avaliação do texto) quanto a reflexão do estudante sobre seu desempenho. No caso da tarefa proposta, entende-se que ela fornece subsídios para compreender em que medida um estudante foi

capaz de apreender o ponto central do texto (a tese da autora). Nesse sentido, pode-se analisar se os resultados a que chegamos com os LLMs revelam algum tipo de contribuição para as tarefas de avaliação de texto, a partir da comparação entre as avaliações humana e automática.

2. Fundamentos

Esta seção apresenta os fundamentos teóricos da investigação, abordando a correção automática de textos, os desafios da avaliação de respostas discursivas curtas, os LLMs e o papel dos *prompts* nas escolhas metodológicas do estudo.

2.1. Correção automática

A escrita é uma habilidade fundamental para os seres humanos que vivem em sociedades letradas. Assim como o seu processo de ensino é complexo, a tarefa de avaliação também o é (Meira et al., 2023). No contexto educacional, as tarefas de escrita estão presentes em todos os níveis de ensino e constituem um dos principais instrumentos de verificação da aprendizagem em diferentes áreas do conhecimento. Mais especificamente na área de língua materna, as avaliações de textos escritos são fortemente marcadas pela subjetividade (Cançado et al., 2020), além de serem uma atividade morosa, exigente e onerosa (Martins & Sousa, 2025). Em razão dessas dificuldades, instituições públicas e privadas têm investido no desenvolvimento e na adoção de sistemas de correção automática, ainda que nem sempre obtenham os resultados esperados (Martins & Sousa, 2025). Esses sistemas partem da premissa de que a competência linguística humana envolvida na avaliação textual pode, ao menos parcialmente, ser transferida para a máquina.

Como se vê em Rassi & Lopes (2023), a correção automática de redações é uma tarefa de processamento de linguagem natural (PLN) em que sistemas computacionais avaliam e/ou atribuem nota a textos em prosa. Nesse caso, um sistema pode implementar soluções distintas, como detectar problemas de escrita, atribuir nota e apresentar um *feedback* ao escritor/estudante. Embora um sistema completo (e complexo) como esse ainda não tenha sido desenvolvido, de acordo com Silveira et al. (2025), já há uma série de testes com ferramentas e modelos que buscam ocorrências de padrões para auxiliar na avaliação. Nesse contexto, Meira et al. (2023, p. 324) afirmam que as ferramentas analíticas de correção automática, se bem integradas às atividades pedagógicas, teriam o potencial de facilitar a tarefa docente, uma vez que com elas é mais fácil "encontrar padrões, lacunas e estratégias de intervenção para apoiar a aprendizagem em diferentes níveis." Logo, o professor teria diagnósticos mais específicos do desempenho dos estudantes, podendo focar naquilo que considera prioritário.

Neste trabalho, focamos na tarefa de atribuição de nota. Rassi & Lopes (2023) afirmam que, assim como ocorre em diferentes sistemas de avaliação humana, a atribuição da nota a um texto pode ser feita globalmente (nota final e única) ou a vários critérios que podem ser somados para se chegar a um total. Modelos de ASAG costumam utilizar aprendizado supervisionado, tomando como base um *corpus* já avaliado por humanos (*dataset*; padrão-ouro). O modelo é treinado com uma parte desses dados e, em seguida, é testado na tarefa proposta. Em nosso trabalho, em razão de um conjunto relativamente pequeno de dados (129 respostas) e, por testarmos a capacidade de LLMs na correção, o direcionamento da tarefa se deu pelo uso de *prompts*. Aqui, inclusive, testamos uma variação de *prompt* para verificar o quanto os modelos flutuam nas respostas (ver Seção 4).

Rassi & Lopes (2023) também sustentam que é essencial que os detalhes do texto solicitado (i.e., redação curta ou longa, por exemplo) e seus critérios de avaliação estejam claros para a tarefa de avaliação automática (assim como deve ser para os humanos). Nesse sentido, as autoras entendem que, ao longo dos anos, as grades de avaliação de textos (os de vestibular e ENEM, mais especificamente) ganharam em publicidade e objetividade de avaliação, o que facilita o treinamento de modelos automáticos — ainda que, como se sabe, o processo subjetivo de correção leve frequentemente os textos a serem corrigidos por diferentes avaliadores, como ocorre no ENEM, em que uma terceira correção no mesmo texto é algo relativamente comum. Como se detalha na Seção 4, nesta pesquisa, ainda que a questão seja discursiva, ela tinha certa objetividade, uma vez que solicitava a ideia central da autora, relacionada a um argumento que ela apresentava no texto. Além disso, a grade de correção apresentava uma resposta esperada e algumas escalas de correção para que os modelos pudessem simular o que fizeram os avaliadores humanos. Esses fatos fizeram com que levantássemos a hipótese de que tanto a avaliação humana quanto a automática apresentariam menor discrepância. Neste trabalho, agregamos essas características da tarefa de ASAG para avançar também em direção à possível contribuição que os LLMs podem dar ao fenômeno.

2.2. Grandes Modelos de Linguagem e *prompts*

Desde sua popularização, os LLMs têm se imposto como uma ferramenta de testes para atividades relacionadas à análise e produção de textos, em virtude de seu poder de interpretar e gerar conteúdos em língua natural, como os humanos (Kasneci et al., 2023). Nesse contexto, as empresas desenvolveram diversos modelos com interfaces amigáveis, tais como o GPT (Radford et al., 2018) da OpenAI, o LLaMA (Touvron et al., 2023a,b) da Meta AI, o Gemini (Gemini Team et al., 2023) do Google DeepMind, o DeepSeek da DeepSeek AI (Liu et al., 2025) e o Sabiá (Abonizio et al., 2024) da Maritaca-AI (i.e., treinado em um grande *corpus* em português brasileiro). O emprego desses modelos, com *prompts* adequados pode contribuir para a qualidade de uma tarefa, tal como a de avaliação de um texto.

Santos (2025) aponta que, entre as vantagens do uso de LLMs para tarefas de ASAG, em vez de tecnologias precedentes, estão o treinamento auto-supervisionado com grande volume de dados e a compreensão semântica mais global desses modelos (com maior ou menor refinamento de *prompts*), permitindo análises qualificadas de respostas não esperadas, mas com sentido. No entanto, conforme também aponta o autor, os LLMs apresentam uma série de inadequações nas saídas das tarefas (incompletude ou irrelevância, por exemplo), o que é problemático para o contexto educacional, em que se espera um maior grau de exatidão em função do caráter avaliativo da escrita, comentado anteriormente.

A forma prototípica de orientar a tarefa de um LLM é por meio de um *prompt*, que pode ser definido como uma instrução ou entrada fornecida a um modelo de linguagem, solicitando uma saída específica (uma resposta ou a continuação de um texto, por exemplo). Assim, um *prompt* orienta o modelo na geração de textos relevantes ou na execução de tarefas específicas (Radford et al., 2018). Segundo os autores, os *prompts* podem ser compostos de questões, descrições de problemas, trechos de um texto ou outras entradas que contribuam para uma resposta coerente por parte do modelo (Radford et al., 2018; Gu et al., 2023). Tomando o aprendizado via *prompts* como essencial, este trabalho variou as instruções, a fim de verificar se o acréscimo ou a subtração de informações traria resultados distintos. Mais especificamente, elaboramos *prompts* com todas as informações possíveis (contexto, papel, texto-base, enunciado da questão, resposta esperada, critérios de avaliação e regras adicionais), com a ausência do texto-base e com a ausência da resposta esperada. Isso nos permitiu levantar as seguintes hipóteses: i) com *prompts* mais completos, os LLMs avaliariam melhor (tendo em vista menor possibilidade de inadequação ou irrelevância); e ii) a avaliação humana ainda seria discrepante do modelo computacional, considerando que ela está sujeita à subjetividade e a considerações mais globais na interpretação do texto (e da resposta esperada). Como esses fatores não são controlados, mantivemos a hipótese de que o comportamento dos modelos pudessem apresentar desempenho inferior ao humano.

3. Trabalhos relacionados

Embora tenha ganhado espaço na área de PLN no Brasil nos últimos anos, a discussão sobre ASAG em português brasileiro ainda é bastante restrita (Rassi & Lopes, 2023; Mello et al., 2024; Martins & Sousa, 2025; Silveira et al., 2025). Nesse cenário, alguns trabalhos ajudam a situar tanto a evolução das abordagens computacionais quanto os desafios ainda associados à confiabilidade da avaliação automática.

Os trabalhos de Galhardi et al. (2018, 2022) se estabelecem como importantes referências de ASAG para o português, fundamentando a avaliação automática em similaridade léxica e na extração de atributos linguísticos. Os autores apresentam bons resultados no uso de modelos computacionais que contribuem para a tarefa de ASAG, sem o envolvimento de LLMs. Embora o presente trabalho compartilhe o propósito de automatizar a correção de respostas curtas para o apoio docente, ele se distancia dessas abordagens ao prescindir da engenharia atributos e de grandes volumes de dados rotulados, utilizando, em vez disso, o raciocínio semântico direto de LLMs. Decidiu-se, entre outras coisas, privilegiar a discussão dos critérios de correção estabelecidos pelo grupo de pesquisa, algo muito mais complexo de se realizar com as técnicas de aprendizado de máquina. Dado o tamanho da amostra, não seria possível um treinamento adequado, capaz de oferecer resultados significativos, ao contrário do que se tem observado com os LLMs. Outro fator a se destacar é que se procurou uma abordagem que pudesse ser replicada, de algum modo, pelos docentes, sem técnicas muito complexas, o que, novamente, parece ser oferecido pelos atuais modelos.

Mais recentemente, com a ascensão dos LLMs, Ye et al. (2025) analisaram a confiabilidade desses modelos como avaliadores automáticos, mostrando que diferentes vieses podem influenciar seus julgamentos.Essas tendências podem expli-

car determinados padrões de comportamento dos modelos e, no caso do presente trabalho, de possíveis atribuições de notas.

Em uma vertente voltada à análise linguística de textos avaliados automaticamente, Meira et al. (2023), utilizando a ferramenta NILC-Metrix (Leal et al., 2024), notaram uma correlação positiva entre notas mais altas e fatores como complexidade sintática e diversidade lexical. Os autores apontam também para a necessidade de mais estudos, com diferentes textos e ferramentas, de modo a enriquecer o campo de análise e buscar soluções inovadoras que contribuam para o desenvolvimento da aprendizagem da escrita (por meio da avaliação). Tendo em vista a emergência e a inovação vinculada aos LLMs, em nossa pesquisa, não analisamos dados linguísticos dos textos, mas avaliamos a capacidade dos modelos de atribuir notas a respostas curtas, em comparação ao padrão-ouro.

Ainda no campo da avaliação automática em português, Martins & Sousa (2025) analisaram três plataformas on-line de correção do ENEM, verificando seu desempenho na atribuição de notas (globais e por critério). No *corpus* com 40 redações, os autores observaram que o desempenho em alguns critérios superou o da nota global. Em razão das limitações observadas, os autores entendem que as plataformas analisadas não substituem a correção humana, sugerindo novas pesquisas, com novos tipos de texto e novas ferramentas. Ainda que seja relativa a um tipo diferente de texto (respostas curtas), a presente pesquisa amplia o conjunto de dados e fornece uma análise integrada da avaliação feita por LLMs.

Analisando o desempenho do GPT-4 na tarefa de ASAG, em comparação a modelos tradicionais, Ferreira Mello et al. (2025) verificaram que o fornecimento de exemplos e o tempo para a resposta foram variáveis de *prompts* que influenciaram o resultado do modelo da OpenAI, confirmando o que a literatura tem observado: a relevância dessas estratégias de *prompt* (*prompt engineering*) para o sucesso de tarefas. Apesar disso, os autores indicam que os modelos tradicionais ainda superaram o GPT-4 na tarefa com os *corpora* analisados por eles.

O trabalho mais próximo ao que desenvolvemos aqui é o de Santos (2025). O autor ressalta a qualidade dos LLMs na interpretação de textos para justificar a escolha dos modelos para a tarefa de ASAG. Em sua pesquisa, Santos (2025) analisa o desempenho de três modelos, combinando suas arquiteturas às variações de *prompt* e aspectos linguísticos dos textos. O autor observou que a estratégia "*few-shot*" (informações básicas da tarefa para o modelo) foi a que mais influenciou positivamente os resultados. Já textos com riqueza lexical média foram os mais desafiadores aos modelos. Em nossa pesquisa, avaliamos cinco LLMs, com variação de *prompt* e sem levar em conta dados linguísticos.

Silveira et al. (2025) apontam que uma dificuldade para se analisar e/ou construir modelos de correção automática está na escassez de grandes *datasets*, sobretudo para o português. Em uma argumentação similar, Rassi & Lopes (2023) entendem que a carência de investigação em português decorre de limitações de recursos, modelos de anotação e bancos de dados. Rassi & Lopes (2023) acrescentam ainda que as avaliações automáticas de textos carecem também de variação de tipo, sendo muito focadas em redações, como as do ENEM. As autoras alertam para a necessidade de se superar esses desafios em prol do aperfeiçoamento sobre o tema. Nesse contexto, este trabalho reforça a necessidade de se desenvolver mais conjuntos de dados padrão-ouro em português brasileiro, com o objetivo de subsidiar pesquisas linguísticas (com ou sem ferramentas computacionais). Especialmente, bancos de dados com textos de estudantes (de todos os níveis) são muito restritos, o que faz com que os trabalhos tenham, também, pouco impacto para a área de ensino.

Assim, ratifica-se que este trabalho se singulariza em três pontos essenciais. Primeiramente, pela comparação entre cinco diferentes modelos de linguagem, o que ainda não se observou na literatura (sobretudo em português brasileiro). Em segundo lugar, pela análise da variação do enriquecimento contextual dos *prompts*, considerando o texto-base (contexto de produção) e os critérios de avaliação (unidade básica para a avaliação e escalonamento das notas). Finalmente, o trabalho agrupa respostas da compreensão da tese de um texto, realizadas por graduandos em Letras, grupo que ainda carece de pesquisas na área linguística.

Nosso foco é a avaliação do desempenho de LLMs na tarefa de ASAG, considerando apenas a atribuição de notas (sem justificativa). Assim, as razões para não solicitar explicações sobre a atribuição da nota são: 1) queríamos atribuir ao LLM a mesma tarefa humana; e 2) nossa intenção é verificar a pertinência da nota global atribuída. De qualquer modo, a fim de testar a influência do detalhamento dos *prompts* para a tarefa, fizemos algumas variações.

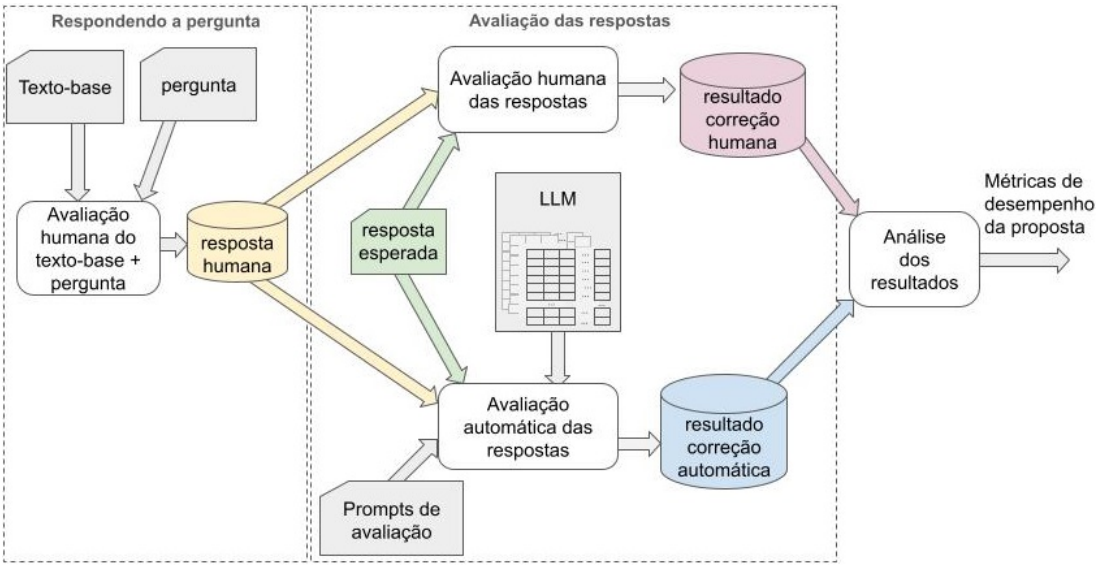


Figura 1: Fluxo de trabalho proposto

4. Metodologia

A Figura 1 ilustra o fluxo de trabalho de avaliação do texto-base e da pergunta discursiva. Ele inicia-se com a apresentação de um texto-base e de uma pergunta discursiva, a partir dos quais o estudante elabora sua resposta. O texto-base usado nos experimentos relatados neste artigo é um texto de opinião, público, retirado de um jornal brasileiro.[1] A resposta do aluno à pergunta constitui o elemento central do processo, pois será submetida a dois tipos distintos de avaliação. Inicialmente, ocorre a avaliação humana, na qual especialistas analisam a resposta considerando o texto-base e o enunciado da questão. O resultado dessa etapa gera a correção humana, que funciona como referência principal (ou padrão ouro) para as comparações posteriores.

Paralelamente, desenvolve-se a avaliação automática por meio de um LLM. Nessa etapa, o sistema recebe como entrada a resposta do estudante, a resposta esperada (gabarito) e um conjunto de *prompts* de avaliação previamente elaborados. Esses *prompts* orientam o LLM a agir de forma semelhante a um avaliador humano, aplicando critérios definidos para atribuir uma pontuação à resposta. Como resultado, obtém-se a correção automática, que representa a interpretação do LLM sobre o desempenho do estudante.

Após a obtenção dos dois tipos de correção, os resultados são integrados em uma etapa de análise comparativa. Nessa fase, busca-se verificar o grau de concordância entre a avaliação humana e a automática, identificando aproximações, divergências e possíveis padrões de erro. Essa comparação é fundamental para compreender o comportamento dos LLMs na tarefa de avaliação de respostas discursivas.

Por fim, a análise gera métricas de desempenho específicas para avaliar a proximidade entre as correções humana e automática, com destaque para a correlação de Pearson, que mede o grau de relação linear entre as avaliações, e para o Kappa ponderado quadrático (QWK, do inglês, *Quadratic Weighted Kappa*), que avalia o grau de concordância entre as pontuações atribuídas, considerando diferentes pesos para as discrepâncias entre as categorias. Complementarmente, são utilizadas métricas de erro, como o erro médio absoluto (MAE, do inglês *Mean Absolute Error*) e a raiz do erro quadrático médio (RMSE, do inglês, *Root Mean Square Error*), que quantificam a diferença entre as pontuações atribuídas pelos humanos e pelo LLM. Essas medidas permitem verificar não apenas se o LLM acompanha a tendência das avaliações humanas e apresenta concordância com elas, mas também o quanto suas pontuações se afastam das avaliações humanas em termos absolutos e quadráticos. Dessa forma, o fluxo de trabalho possibilita uma análise mais precisa do desempenho da correção automática e de seu potencial como ferramenta de apoio à avaliação da compreensão textual.

4.1. Coleta de dados

A coleta de dados para esta pesquisa foi realizada nos laboratórios de informática da Universidade Tecnológica Federal do Paraná (cursos de

[1]Disponível em: https://github.com/osmarbraz/aqdc_lmm/blob/main/texto_versao_digital.pdf

Letras), com 134 estudantes voluntários de graduação em Letras, do primeiro ao oitavo semestre do curso, entre dezembro de 2023 e abril de 2024. A maioria deles vinha de escola pública (99) e dividia-se em diferentes faixas etárias, com predomínio dos mais jovens (74 alunos de 18 a 20 anos; 49 de 21 a 28 anos; e 11 acima de 29 anos). Além disso, 104 estudantes (78% do total) cursavam sua primeira graduação. A intenção da pesquisa era, inicialmente, verificar se as anotações interferiam no desempenho da leitura dos estudantes.[2] A tarefa relevante para a presente pesquisa consistia na leitura de um texto de opinião (324 palavras) em formato digital e posterior resposta a quatro questões (três objetivas e uma discursiva). Na questão discursiva, os estudantes deveriam completar o enunciado "Complete a frase a seguir: No texto, a autora usa o GPS como um exemplo para defender a ideia de que. . . ", formulando, com suas próprias palavras, a tese central do texto de opinião lido. Com isso, buscou-se avaliar sua compreensão global do texto e sua capacidade de síntese interpretativa.

Ao término da etapa de coleta de dados e da avaliação dos instrumentos de pesquisa, foram obtidas 134 respostas válidas iniciais. Em seguida, realizou-se o processo de tratamento e depuração dos dados, que consistiu na exclusão de registros com itens não respondidos. Após essa etapa, a amostra final foi constituída por 129 respostas válidas, as quais foram consideradas nas análises estatísticas desta pesquisa.[3]

4.2. LLMs

Para investigar o desempenho na tarefa ASAG em português, foram selecionados cinco LLMs cujos perfis se complementam tanto em arquitetura quanto em contexto de uso, abrangendo soluções consolidadas, alternativas nacionais e modelos recentes de alto desempenho:

- **ChatGPT 5.3** (OpenAI)[4]: essa versão mantém o núcleo de raciocínio avançado das gerações anteriores, incluindo capacidade multimodal, ampla janela de contexto e melhorias em consistência e precisão, aliadas a maior eficiência computacional (Brown et al., 2020).
- **Sabiá 4 2026-01-06** (Maritaca-AI)[5]: representa a vertente nacional, sendo um modelo treinado e ajustado com forte ênfase em português brasileiro. Essa especialização permite melhor compreensão de nuances linguísticas, variações regionais e aspectos culturais que influenciam a avaliação de respostas discursivas, além de apresentar custo competitivo por token e maior adequação ao contexto educacional brasileiro (Laitz et al., 2026).
- **Gemini 3** (Google DeepMind)[6], incluído como alternativa eficiente para cenários que exigem escalabilidade. O modelo combina uma ampla janela de contexto com inferência rápida e otimizada, mantendo bom desempenho em tarefas de compreensão textual e avaliação semântica (Mughal et al., 2026).
- **LLaMA-3.3 70B Instruct** (Meta AI)[7]: incorporado por seu forte desempenho em tarefas de raciocínio e geração de texto, especialmente em configurações *instruction-tuned*. O modelo apresenta alta capacidade de generalização e bom equilíbrio entre qualidade e custo quando executado em infraestrutura própria (Dsouza, 2025; Abdullah et al., 2025).
- **DeepSeek-V3.2** (DeepSeek-AI)[8]: incluído como representante de uma nova geração de modelos otimizados para eficiência e raciocínio estruturado. Destaca-se por seu desempenho competitivo em *benchmarks* de compreensão e geração textual (Neha & Bhati, 2025; Liu et al., 2025).

4.3. Prompts

O *prompt* base, utilizado em uma abordagem do tipo *zero-shot* — isto é, sem fornecimento de exemplos prévios ao modelo —, é empregado para solicitar ao LLM a avaliação da questão e é transcrito a seguir. Inicialmente, o *prompt* define o contexto, indicando que se trata da correção de uma atividade acadêmica. Em seguida, determina o papel da inteligência artificial (IA), que deve agir como um professor universitário, ou seja, realizar uma avaliação criteriosa, imparcial e baseada em critérios definidos. O *prompt* também organiza as informações de entrada por meio de marcadores específicos: o texto-base pelo marcador "`####`", o enunciado da questão pelo marcador "`###`" e a resposta do aluno pelo marcador "`##`" são apresentados em blocos delimitados. Isso facilita a identificação de cada elemento necessário para a correção e garante que

[2] Os resultados gerais dessa investigação podem ser vistos em Nunes & Bertucci (2025); Busato (2025); Lima (2025).

[3] O texto-base e os dados da pesquisa estão disponíveis em: `https://github.com/osmarbraz/aqdc_lmm/`

[4] `https://www.chatgpt.com/`

[5] `https://www.maritaca.ai/`

[6] `https://gemini.google.com/`

[7] `https://huggingface.co/chat/conversation/`

[8] `https://chat.deepseek.com/`

a análise seja feita com base em todas as partes relevantes da atividade. Outro ponto central é a apresentação da resposta esperada, que orienta o LLM sobre a ideia principal que deve ser identificada na resposta do aluno — no caso, a relação entre IA (ou algoritmos) e as escolhas humanas. Essa referência funciona como um parâmetro para avaliar o nível de adequação da resposta. Além disso, o *prompt* inclui critérios de avaliação detalhada, com pontuações distribuídas em níveis (0 a 40 pontos). Cada faixa de pontuação descreve o grau de correspondência entre a resposta do aluno e a tese do texto, permitindo uma classificação mais precisa. Também são previstas pontuações intermediárias, o que torna a avaliação mais flexível. As regras adicionais reforçam critérios importantes, como desconsiderar erros gramaticais, valorizar respostas concisas que expressem a ideia central e limitar a pontuação de respostas que apenas copiam o texto-base ou não completam o enunciado. O *prompt* ainda orienta um procedimento interno de avaliação, indicando que o LLM deve comparar a resposta do aluno com o texto-base e a resposta esperada, classificar conforme os critérios e ajustar a nota, mas sem expor esse raciocínio no resultado final. Por fim, define-se a saída esperada, que deve conter apenas a nota final (em valores pré-estabelecidos), sem qualquer explicação adicional, o que garante padronização e objetividade na resposta.

```
Contexto: Você está corrigindo uma atividade de
interpretação textual de um curso de ensino
superior.

Papel: Assuma o papel de um professor
universitário avaliando objetivamente a resposta
de um estudante.

O texto-base está entre os marcadores ####:
####
<TEXTO_BASE>
####

O enunciado da questão está entre os marcadores
###:
###
Complete a frase a seguir: NO TEXTO, A AUTORA
USA O GPS COMO UM EXEMPLO PARA DEFENDER A IDEIA
DE QUE...
###

A resposta do aluno está entre os marcadores ##:
##
<RESPOSTA_ALUNO>
##

Resposta esperada: ... a IA / os algoritmos podem
auxiliar a realizar escolhas (e usar a
inteligência/cabeça para outras coisas); ou ...
a IA não é uma ferramenta de escolha no lugar
do ser humano.

Critérios de avaliação:
40 Pontos: O texto apresenta a tese da autora,
relacionada com a contribuição da IA/algoritmos
para as escolhas humanas.
30 Pontos: A resposta indica que o enfoque da
autora é tratar sobre IA e escolhas humanas,
mas desenvolve apenas um dos fatores.
20 Pontos: Texto foca no GPS, sem relacionar com
IA ou escolhas humanas; respostas que sejam
apenas citações não devem passar dessa nota.
10 Pontos: Não há relação entre a tese e o
argumento, ainda que o texto apresente resquícios
de IA, algoritmo ou escolha; textos que não
continuam o enunciado podem não passar desta
nota.
0 Pontos: Não tratar nem de IA, nem de
algoritmos, nem de escolha.

Pontuações intermediárias (5, 15, 25, 35) podem
ser atribuídas quando a resposta estiver entre
dois níveis da escala.

Regras adicionais:
- Erros gramaticais não devem influenciar a nota.
- Respostas muito curtas podem ser válidas se
expressarem corretamente a ideia central.
- Respostas que apenas copiam trechos do
texto-base não devem receber mais que 20 pontos.
- Respostas que não completam o enunciado não
devem receber mais que 10 pontos.

Procedimento de avaliação (realize o raciocínio
internamente e não o apresente):
1. Compare a resposta do aluno com a resposta
esperada e com o texto-base.
2. Classifique a resposta segundo os critérios
da escala.
3. Ajuste a pontuação se necessário usando
valores intermediários.

Saída esperada:
Retorne apenas a nota final entre os valores
possíveis: 0, 5, 10, 15, 20, 25, 30, 35 ou 40

Não inclua explicações ou qualquer outro texto.
```

Utilizamos três versões deste *prompt*: a primeira inclui o texto-base e a resposta esperada; a segunda, o texto-base, mas sem a resposta esperada; e a terceira e última sem o texto-base e com a resposta esperada. Optamos por não incluir uma versão do *prompt* sem o texto-base e sem a resposta esperada porque isso ocasionaria a falta do contexto mínimo necessário para realizar a tarefa de avaliação.

4.4. Considerações éticas e tratamento dos dados

Esta pesquisa foi submetida à apreciação e aprovada pelo Comitê de Ética em Pesquisa na instituição de origem (registro CAAE 88328218.6.0000.5547). Ela contou com a participação voluntária de 129 estudantes do curso de Letras da Universidade Tecnológica Federal do

Paraná. Os critérios de inclusão foram: ser licenciando em Letras, ter idade acima de 18 anos e ter assinado a autorização para uso dos dados. Antes das análises, as respostas foram tratadas de modo a preservar a identidade dos participantes, com remoção de informações que pudessem permitir identificação direta ou indireta dos participantes. Assim, o conjunto utilizado nas avaliações humanas e automáticas foi composto apenas pelas respostas textuais necessárias à tarefa, sem associação a nomes, matrículas, e-mails ou outros identificadores pessoais.

O tratamento dos dados observou os princípios da Lei Geral de Proteção de Dados Pessoais (LGPD), especialmente no que se refere à finalidade, necessidade, minimização e segurança no uso de informações estudantis. Considerando que parte dos experimentos envolveu o envio de respostas a ferramentas comerciais de LLM, adotou-se como medida de mitigação o envio exclusivo de dados anonimizados e estritamente necessários à avaliação automática. Ainda assim, reconhece-se que o uso de plataformas comerciais implica riscos específicos relacionados ao processamento externo dos dados, às políticas de retenção e ao possível uso das interações pelos provedores. Por esse motivo, os resultados devem ser interpretados também à luz dessas limitações éticas e operacionais, reforçando a necessidade de cautela no uso de LLMs comerciais em contextos educacionais com dados de estudantes.

4.5. Métricas de comparação

A avaliação dos experimentos foi estruturada por meio de quatro métricas complementares: o coeficiente de correlação de Pearson, QWK, MAE e a RMSE. Essa combinação permite uma análise balanceada, contemplando tanto a relação linear entre as variáveis quanto a precisão quantitativa das predições.

O coeficiente de correlação de Pearson quantifica o grau de associação linear entre as notas geradas pelos LLMs e aquelas atribuídas por avaliadores humanos, indicando o quanto as variações em uma medida acompanham as variações na outra e fornecendo uma medida robusta de alinhamento entre os valores (Pearson, 1895). O QWK é uma métrica que avalia a concordância entre classificações ordinais, corrigindo o efeito do acaso e penalizando mais severamente erros entre categorias mais distantes (Cohen, 1968). Para quantificar a incerteza associada à estimativa, foram calculados intervalos de confiança de 95% (IC95%) por *bootstrap* não paramétrico com 10.000 reamostragens, utilizando os percentis 2,5% e 97,5% como limites do intervalo. O MAE expressa a discrepância média absoluta entre a nota prevista e a nota de referência, oferecendo uma interpretação direta do erro médio (Willmott & Matsuura, 2005). Já o RMSE enfatiza erros maiores ao elevar ao quadrado as diferenças individuais, sendo útil para identificar o impacto de desvios extremos no desempenho do modelo (Willmott & Matsuura, 2005).

5. Experimentos e resultados

Nesta seção, adicionamos detalhes aos dois experimentos e apresentamos os seus resultados. Ao final, analisamos e discutimos os dados.

5.1. Avaliação humana

As respostas da questão discursiva curta obtidas no experimento foram avaliadas por dois corretores humanos, os quais atribuíram notas de 0 a 40, seguindo os critérios estipulados para a correção (os mesmos presentes nos *prompts* apresentados anteriormente). Esses avaliadores eram formados na área de Letras e discentes do curso de mestrado em Estudos de Linguagens da UTFPR. Vale ressaltar que os critérios foram discutidos e refinados por todos os membros da equipe de pesquisa, ainda que apenas os mestrandos tenham feito a correção das questões. Além disso, antes da avaliação, houve um treinamento com um conjunto de 20 respostas, avaliadas por todo o grupo, em que possíveis discrepâncias foram discutidas.

A Tabela 1 mostra que os avaliadores humanos apresentam uma correlação linear de Pearson forte (igual a 0,91), sugerindo pouca discrepância tanto entre si quanto com a média geral, confirmando a hipótese inicialmente levantada.

Avaliadores	*Avaliador 1*	*Avaliador 2*	*Média*
Avaliador 1	-	0,91	0,98
Avaliador 2	0,91	-	0,98
Média	0,98	0,98	-

Tabela 1: Matriz de correlação de Pearson das notas atribuídas às questões pelos avaliadores humanos e sua média

A Tabela 2 mostra que os avaliadores humanos também apresentam alta concordância pelo QWK, com valor de 0,91 entre si e IC95% entre 0,87 e 0,94. Além disso, a concordância de cada avaliador com a média geral também foi elevada (0,97 para o Avaliador 1, com IC95% entre 0,96 e 0,98; e 0,98 para o Avaliador 2, com IC95% entre 0,97 e 0,98). O fato de esses intervalos serem

Avaliadores	*Avaliador 1*	*Avaliador 2*	*Média*
Avaliador 1	—	0,91 (IC95%: 0,87–0,94)	0,97 (IC95%: 0,96–0,98)
Avaliador 2	0,91 (IC95%: 0,87–0,94)	—	0,98 (IC95%: 0,97–0,98)
Média	0,97 (IC95%: 0,96–0,98)	0,98 (IC95%: 0,97–0,98)	—

Tabela 2: Kappa Ponderado Quadrático (QWK) das notas atribuídas às questões pelos avaliadores humanos e sua média

estreitos reforça a consistência das estimativas e sugere pouca discrepância nas avaliações, confirmando a hipótese inicialmente levantada. Para o cálculo dessa métrica, as notas fracionadas foram arredondadas, o que ocorreu em apenas uma nota do Avaliador 1, em nenhuma nota do Avaliador 2 e em 36 notas da média.

5.2. Avaliação automática

Os experimentos foram realizados com cinco LLMs entre os dias 05/03/2026 e 23/05/2026 via *chat*. Cada *prompt* foi submetido uma única vez a cada LLM, com reinício do *chat* a cada nova submissão para garantir independência nas respostas. Para estes LLMs, usamos as configurações padrão de hiperparâmetros, incluindo *temperature* e *top_p*. Para todos os *prompts* submetidos, os LLMs retornaram somente a nota final, sem texto adicional. Experimentos com amostras revelaram não haver mudanças significativas na nota atribuída por cada LLM em repetidas chamadas com a mesma requisição (mesma resposta de aluno), donde preferimos usar os nossos recursos limitados para processar mais dados distintos com mais LLMs.

A Tabela 3 sintetiza o desempenho de diferentes LLMs sob variações na composição do *prompt*, considerando a presença ("`+`") ou ausência ("`-`") dos componentes texto-base ("`T`") e resposta esperada ("`R`"). Os resultados foram avaliados por meio de quatro métricas: correlação de Pearson, Kappa ponderado quadrático (QWK) com intervalo de confiança de 95% (IC95%), erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE). Para a métrica QWK, notas com número fracionário foram arrendondadas. Os valores em negrito indicam o melhor desempenho dentro de cada combinação de componentes de *prompt* e LLM.

De modo geral, observa-se que o LLM Gemini 3 apresentou o melhor desempenho global. A configuração "`+T+R`" alcançou a mais alta correlação de Pearson (0,68), além de apresentar o menor MAE (8,61), indicando maior proximidade entre as predições do modelo e as notas médias dos avaliadores. Essa tendência também se confirma pelo QWK: o Gemini 3 obteve os maiores valores da métrica nas configurações "`+T+R`" e "`+T-R`" (ambas com 0,60), com IC95% semelhantes (0,49–0,69 e 0,48–0,69, respectivamente), sinalizando maior concordância ponderada com a avaliação humana. Ainda assim, como esses intervalos se sobrepõem aos de outros resultados competitivos, como o GPT-5.3 "`+T-R`" (QWK de 0,53; IC95%: 0,40–0,62) e o Sabiá-4 "`-T+R`" (QWK de 0,50; IC95%: 0,39–0,59), a superioridade do Gemini deve ser interpretada com cautela. Já o menor RMSE (10,87) foi obtido pela variação Sabiá-4 "`-T+R`", que também apresentou QWK de 0,50, sugerindo maior estabilidade nas previsões mesmo com a ausência do componente "T". O LLM GPT-5.3 apresentou desempenho consistente, com destaque para a configuração "`+T-R`", que obteve correlação de 0,60 e QWK de 0,53, além de valores relativamente baixos de MAE (8,86) e RMSE (11,45). Esses resultados indicam que, para esse modelo, a remoção do componente "R" pode contribuir positivamente para o desempenho. O LLaMA-3.3 apresentou desempenho intermediário, com melhores resultados na configuração "`-T+R`", que alcançou correlação de 0,52, QWK de 0,46 (IC95%: 0,34–0,56) e reduções nos erros (MAE de 9,07 e RMSE de 11,24). O LLM DeepSeek-V3.2 demonstrou resultados moderados, sendo "`+T+R`"sua melhor configuração na correlação de Pearson, com 0,49, e também seu maior QWK (0,40; IC95%: 0,27–0,52). Entretanto, apresentou valores de erro relativamente elevados em comparação com outros modelos, indicando menor precisão geral.

Por fim, o Sabiá-4 apresentou os menores valores de QWK nas configurações com texto-base (0,32 em "`+T+R`" e 0,31 em "`+T-R`"), com IC95% entre 0,21–0,42 e 0,20–0,42, respectivamente. Ainda assim, a configuração "`-T+R`" obteve a maior correlação de Pearson (0,63) dentro do próprio modelo, os menores erros (MAE de 8,90 e RMSE de 10,87) e o maior QWK do Sabiá-4 (0,50; IC95%: 0,39–0,59), sugerindo que, apesar da baixa taxa de acerto, o modelo consegue manter certa consistência nas previsões e concordância ponderada moderada com a avaliação humana. Em síntese, os resultados indi-

LLMs	Pearson	QWK	MAE	RMSE
LLaMA-3.3 `+T+R`	0,45	0,37 (IC95%: 0,24–0,48)	9,79	12,11
LLaMA-3.3 `+T-R`	0,47	0,34 (IC95%: 0,22–0,44)	10,10	12,50
LLaMA-3.3 `-T+R`	0,52	0,46 (IC95%: 0,34–0,56)	9,07	11,24
GPT-5.3 `+T+R`	0,53	0,46 (IC95%: 0,33–0,56)	9,77	12,68
GPT-5.3 `+T-R`	0,60	0,53 (IC95%: 0,40–0,62)	8,86	11,45
GPT-5.3 `-T+R`	0,55	0,47 (IC95%: 0,35–0,55)	9,89	12,54
Gemini 3 `+T+R`	**0,68**	**0,60** (IC95%: 0,49–0,69)	**8,61**	11,17
Gemini 3 `+T-R`	0,65	**0,60** (IC95%: 0,48–0,69)	8,73	11,09
Gemini 3 `-T+R`	0,37	0,38 (IC95%: 0,21–0,52)	11,21	13,95
Sabiá-4 `+T+R`	0,51	0,32 (IC95%: 0,21–0,42)	10,63	13,49
Sabiá-4 `+T-R`	0,49	0,31 (IC95%: 0,20–0,42)	10,87	13,61
Sabiá-4 `-T+R`	0,63	0,50 (IC95%: 0,39–0,59)	8,90	**10,87**
DeepSeek-V3.2 `+T+R`	0,49	0,40 (IC95%: 0,27–0,52)	10,43	13,56
DeepSeek-V3.2 `+T-R`	0,44	0,34 (IC95%: 0,21–0,45)	11,15	14,45
DeepSeek-V3.2 `-T+R`	0,39	0,34 (IC95%: 0,21–0,45)	10,05	12,03

Tabela 3: Desempenho de LLMs com diferentes combinações de componentes no *prompt* em relação à nota média atribuída às questões pelos avaliadores humanos

cam que o impacto dos componentes do *prompt* varia conforme o modelo, não havendo uma configuração universalmente ótima. A análise do QWK, considerando também os IC95%, reforça essa interpretação: embora o Gemini 3 tenha mantido os maiores valores pontuais de QWK em "`+T+R`" e "`+T-R`" (0,60), a sobreposição dos intervalos com GPT-5.3, LLaMA-3.3 e Sabiá-4 recomenda cautela ao hierarquizar os modelos apenas por essa métrica. Ainda assim, a combinação "`+T+R`" tende a favorecer modelos mais robustos, como o Gemini 3, enquanto a remoção de um dos componentes pode beneficiar LLMs específicos, como observado no GPT-5.3, no LLaMA-3.3 e no Sabiá-4.

A Tabela 4 apresenta uma comparação detalhada das estatísticas descritivas das notas obtidas por avaliadores humanos e por LLMs com diferentes combinações de *prompts*. O *prompt* contendo o texto-base e a resposta esperada é indicado por "`+T+R`"; aquele com o texto-base e sem a resposta esperada, por "`+T-R`"; e aquele sem o texto-base e com a resposta esperada, por "`-T+R`". As métricas analisadas incluem média, desvio padrão, moda, valor máximo e valor mínimo, permitindo uma análise aprofundada do desempenho de cada LLM e avaliador.

A Tabela 4 evidencia que os avaliadores humanos apresentam médias mais baixas ($\approx 21,9$) e alta variabilidade nas notas, utilizando toda a escala de avaliação (0 a 40). Em contraste, os LLMs atribuem, em geral, notas mais altas (entre 25 e 30) e com menor dispersão, indicando maior consistência, porém menor uso dos extremos. Entre os modelos, Sabiá-4 e DeepSeek-V3.2 apresentam as maiores médias, enquanto GPT-5.3 e Gemini 3 mantêm valores intermediários, e LLaMA-3.3 apresenta médias mais próximas das humanas. De modo geral, os LLMs tendem a concentrar suas avaliações em valores intermediários e evitar notas muito baixas. Esses resultados indicam que, em comparação com humanos, os modelos apresentam um comportamento mais homogêneo e um viés de atribuição de notas mais elevadas.

5.3. Discussão dos resultados

Com relação aos dados sobre a avaliação humana, havíamos levantado a hipótese de que haveria menor discrepância entre os avaliadores, em razão da objetividade da questão, da existência de uma escala de notas (critérios) e de um treinamento anterior à correção de todos os dados (ver Subseção 2.1). De fato, os dados (ver Tabela 1) mostram que os avaliadores tiveram uma alta taxa de concordância entre si e com a média geral, reforçando que efeitos de subjetividade podem ser atenuados no processo de avaliação Cançado et al. (2020); Rassi & Lopes (2023).

Ainda focando na objetividade, sugerimos que isso poderia oferecer resultados melhores. Conforme se vê na Tabela 3, a objetividade parece não ter influenciado na discrepância, visto que houve uma grande variação da correlação de Pearson entre os modelos: veja, por exemplo, a diferença entre o LLaMA (0,45) e o Gemini (0,68) na variação "`+T+R`" e entre o DeepSeek (0,39) e

Avaliadores e LLMs	Média	Desvio Padrão	Moda	Máximo	Mínimo
Avaliador 1	21,79	13,16	40,00	40,00	0,00
Avaliador 2	22,02	13,56	40,00	40,00	0,00
Média	21,91	13,05	40,00	40,00	0,00
LLaMA-3.3 `+T+R`	25,19	6,39	30,00	40,00	10,00
LLaMA-3.3 `+T-R`	26,71	5,76	30,00	40,00	10,00
LLaMA-3.3 `-T+R`	23,53	6,83	20,00	40,00	10,00
GPT-5.3 `+T+R`	28,02	8,20	25,00	40,00	10,00
GPT-5.3 `+T-R`	26,63	7,85	30,00	40,00	10,00
GPT-5.3 `-T+R`	27,98	8,13	30,00	40,00	10,00
Gemini 3 `+T+R`	27,75	9,28	25,00	40,00	10,00
Gemini 3 `+T-R`	26,67	9,59	35,00	40,00	5,00
Gemini 3 `-T+R`	22,13	11,68	10,00	40,00	10,00
Sabiá-4 `+T+R`	29,34	5,58	30,00	40,00	15,00
Sabiá-4 `+T-R`	29,38	5,81	35,00	40,00	15,00
Sabiá-4 `-T+R`	25,43	6,62	25,00	35,00	10,00
DeepSeek-V3.2 `+T+R`	28,95	8,46	20,00	40,00	15,00
DeepSeek-V3.2 `+T-R`	29,96	8,34	20,00	40,00	15,00
DeepSeek-V3.2 `-T+R`	22,44	5,94	20,00	40,00	10,00

Tabela 4: Estatísticas das notas atribuídas por humanos e LLMs

o Sabiá (0,63), na variação "`-T+R`". A métrica QWK reforça essa leitura, pois também apresenta diferenças relevantes entre modelos e configurações: na variação "`+T+R`", por exemplo, o Gemini alcançou QWK de 0,60, enquanto o Sabiá obteve 0,32; já na variação "`-T+R`", o Sabiá chegou a 0,50, ao passo que DeepSeek e Gemini registraram 0,34 e 0,38, respectivamente. Assim, embora a formulação objetiva da tarefa e dos critérios tenha favorecido a avaliação humana, os resultados automáticos indicam que o grau de concordância dos LLMs com os avaliadores humanos continuou sensível ao modelo e à configuração do *prompt*.

Quanto ao detalhamento dos *prompts*, algo discutido, por exemplo, em Ferreira Mello et al. (2025) e Santos (2025), levantamos a hipótese de que a inclusão do texto-base ("`T`"), dos critérios detalhados e da resposta esperada ("`R`") contribuiria para uma avaliação mais próxima à humana. As métricas, porém, mostraram que a ausência de "`T`" influenciou negativamente o Gemini e o DeepSeek, enquanto beneficiou o LLaMA e o Sabiá. Quando se observam as notas médias (Tabela 4), vê-se que a ausência de "`T`" fez com que a média de notas caísse em quase todos os modelos (menos no GPT), além de reduzir a moda (nota mais frequente) na maioria dos LLMs. No geral, porém, deve-se ressaltar que a presença do texto foi mais importante do que a resposta esperada indicada explicitamente no *prompt*. Atribuímos isso à presença dos critérios, que detalham a escala de notas (a resposta esperada está presente nos critérios da nota máxima).

Reforça-se que os principais desvios dos LLMs foram em relação às notas extremas (zero e 40). No entanto, é preciso destacar que, em nenhum caso referente as três configurações, houve a atribuição da nota mínima (zero) pelos modelos, ao contrário da nota máxima, que ocorreu algumas vezes (destaque para o GPT com 23 atribuições, seguido do Gemini, com 16). Diante desse cenário, destacamos que as maiores discrepâncias (MAE) dos modelos foram justamente em relação às notas zero: no geral, enquanto os avaliadores humanos se sentiram à vontade para atribuir essa nota aos textos, os modelos divergiram grosseiramente. Quando se observam as respostas, percebe-se que os LLMs tendem a encontrar algum nexo mínimo de sentido, seja com relação ao texto-base, seja em relação aos critérios de correção ou à resposta esperada. Por isso, uma possível explicação para a divergência, sobretudo na configuração "`+T+R`", é que os modelos devem ter priorizado o texto-base, procurando, nas respostas, alguma relação com essa fonte. A Tabela 5 mostra alguns casos de discrepância, retirados do conjunto de dados. Além das notas atribuídas, ela apresenta o ID do informante e a resposta do aluno.

Ainda nesse ponto, diferentes autores têm sublinhado o papel da avaliação como parte do processo de aprendizagem da escrita: mesmo em respostas ruins, é necessário se destacar o esforço

ID	Resposta	Média humana	*LLaMA 3.3*	*GPT 5.3*	*Gemini 3*	*Sabiá 4*	*DeepSeek V3.2*
1	Esse recurso deve auxiliar a pessoa que o está usando, enquanto este pode focar em outras coisas além de planejar rota.	0	30	25	30	35	40
11	É um meio rápido e simplificado de solucionar problemas, poupando o tempo do usuário, que tem outras coisas com que se preocupar.	0	30	25	25	25	20
68	Ele calcula as melhores rotas, com base em nossas preferências pré-selecionadas, enquanto colocamos nossos esforços em outras atividades.	10	30	40	20	35	30
132	São legítimos calculadores inteligentes de rotas.	0	20	20	20	20	20

Tabela 5: Exemplos de discrepâncias em respostas com notas humanas baixas

do aluno para ajudá-lo formação, sem prioridade à nota (Bazarim et al., 2023; Marcuschi, 2004; Ruiz, 2010). Nesse aspecto, percebe-se que os LLMs foram mais cuidadosos que os humanos ao evitarem notas zero (ainda que divergindo dos critérios indicados).

De volta aos dados, eles sugerem que grandes discrepâncias ocorreram com textos que não têm grandes problemas de coesão ou coerência, nem são ambíguos. Percebemos que textos que receberam uma nota generosa dos modelos — em relação às notas humanas — são semelhantes ao texto-base, garantindo alguma similaridade semântica, fator que não foi considerado na avaliação humana, dada a tarefa proposta e os critérios de correção. A resposta 132, por exemplo, é uma cópia de um trecho do terceiro parágrafo do texto-base. Esse fato fez com que os avaliadores humanos desconsiderassem a qualidade da resposta, enquanto todos os modelos atribuíram 50% do total da nota (20/40). Assim, uma possível explicação é justamente o fato de os modelos trabalharem com noções de similaridade entre a resposta e o contexto (texto-base, critérios e resposta esperada) para atribuir notas melhores. Nossa hipótese explicativa é corroborada por outro dado. Se a prevalência de similaridade com o texto base, espera-se que respostas mais genéricas (em relação ao texto-base) sejam avaliadas com peso menor pelos modelos. A Tabela 6 mostra alguns exemplos de textos bem avaliados que receberam uma avaliação menor pelos modelos.

Seguindo a argumentação de Ye et al. (2025), uma possibilidade explicativa seria de que algum viés estaria gerando o padrão de notas dos LLMs. Nesse aspecto, observamos, por exemplo, o critério de "verborragia", uma tendência dos modelos a julgarem como melhores os textos com maior número de palavras.[9] Os dados recortados não mostram uma prevalência categórica desse critério: modelos como Deepseek v3.2, GPT 5.3 e Gemini 3 se mostraram mais resistentes a atribuir maior nota em função do número de palavras, enquanto o LLaMA e o Sabiá-4 parecem considerar mais esse fator.

Acrescenta-se que, se o peso do texto-base é grande para a análise dos LLMs, espera-se que na configuração "`-T+R`", os dados sejam mais próximos da avaliação humana, sobretudo nos casos de nota mais baixa. A Tabela 7 apresenta os resultados comparativos entre a configuração "`+T+R`" (*prompt* com texto-base e resposta esperada) e a configuração "`-T+R`" (apenas a resposta esperada).

Os dados revelam que há duas alternativas nesses casos: ou os modelos mantêm a nota ou a diminuem. Isso, novamente, revela que alguns modelos parecem mais sensíveis à presença do texto-base para a avaliação da resposta, como indicam os casos do LLaMA e do Sabiá. Quando se observa a Tabela 3, percebe-se que esses dois modelos, tiveram melhores índices de Pearson, QWK, MAE e RMSE melhores na configuração "`-T+R`".

Concluímos, assim, que, mesmo diante dos critérios e de uma resposta esperada, alguns modelos parecem ter privilegiado a relação da resposta do estudante com o texto-base, fato que pode estar relacionado com uma exposição prévia dessa fonte no treinamento dos LLMs. No entanto, diferentemente de pesquisas que solicitam aos LLMs a interpretação direta de textos, neste trabalho os modelos foram utilizados para avaliar respostas únicas produzidas por participantes humanos após uma atividade de leitura. Esse fato ameniza a interferência que um possível treinamento anterior teria no desempenho dos LLMs.

[9] No original, "verbosity"; ocorre quando um "LLM judges favor longer responses, even if they are not as clear, high-quality, or accurate as shorter alternatives." (Ye et al., 2025, p. 4).

ID	Resposta	Média humana	*LLaMA 3.3*	*GPT 5.3*	*Gemini 3*	*Sabiá 4*	*DeepSeek V3.2*
31	As ferramentas tecnológicas permeiam as diversas esferas do nosso cotidiano, nos auxiliando nas atividades do dia a dia.	32,5	10	15	10	20	20
104	IAs são presentes na sociedade a décadas, terceirizando tarefas simples para que o usúario possa se concentrar em outros problemas.	40	30	30	40	25	20
115	A inteligência artificial, já é presente no dia a dia, e pode ajudar a resolver coisas simples, que normalmente demandam muito mais tempo quando o ser humano precisa fazer.	32,5	30	25	25	30	20

Tabela 6: Exemplos de discrepâncias em respostas com notas humanas altas

ID	Média humana	*LLaMA 3.3*		*GPT 5.3*		*Gemini 3*		*Sabiá 4*		*DeepSeek V3.2*	
		+T+R	−T+R	+T+R	−T+R	+T+R	−T+R	+T+R	−T+R	+T+R	−T+R
1	0	30	20	25	25	30	10	35	25	40	20
11	0	30	20	25	25	25	25	25	25	20	20
31	32,5	10	10	15	10	10	10	20	15	20	20
68	10	30	20	40	30	20	20	35	30	30	20
104	40	30	10	30	30	40	10	25	25	20	20
115	32,5	30	20	25	25	25	15	30	25	20	20
132	0	20	20	20	20	20	10	20	10	20	20

Tabela 7: Comparação das discrepâncias por configuração

A última hipótese dizia respeito à discrepância entre humanos e LLMs, em virtude da subjetividade. Desse ponto de vista, entendemos que a subjetividade humana continuou sendo um ponto importante, uma vez que os avaliadores utilizaram de forma mais ampla a escala de notas. Já os LLMs atribuíram notas mais altas (ver Tabela 4) e evitaram dar notas baixas (a nota mínima foi 5, enquanto a dos humanos foi 0). Esses fatos reforçam a maior sensibilidade que os avaliadores humanos empregam na atribuição de notas, identificando desvios cruciais de interpretação (do texto ou do enunciado), o que parece difícil aos LLMs. Assim, além dos dados gerais, esses pontos específicos indicam desempenho inferior dos modelos em relação à avaliação humana. Nossos dados reforçam a conclusão de Martins & Sousa (2025) de que, apesar de resultados melhores, a tarefa de ASAG ainda representa um desafio para os sistemas atuais.

6. Conclusão e trabalhos futuros

Este trabalho insere-se no contexto das discussões contemporâneas sobre o uso de tecnologias emergentes na educação, com ênfase nos LLMs aplicados à avaliação de respostas discursivas curtas. Considerando os desafios históricos relacionados à subjetividade, ao custo e ao tempo da avaliação humana, a pesquisa buscou investigar o potencial dessas tecnologias como ferramentas de apoio à análise textual em ambientes educacionais, especialmente no cenário brasileiro, ainda carente de estudos e recursos voltados à tarefa de ASAG em língua portuguesa.

Como principal contribuição metodológica, o estudo propõe uma comparação entre a avaliação humana — tomada como padrão de referência — e as capacidades automatizadas dos LLMs, explorando diferentes variações de *prompts* em um cenário *zero-shot*. A originalidade do método reside tanto na comparação sistemática entre múltiplos modelos de linguagem quanto na análise do impacto da estrutura dos *prompts* (presença ou ausência de texto-base e resposta esperada) no desempenho da avaliação automática. Essa estratégia permitiu investigar não apenas a capacidade dos modelos, mas também a sensibilidade da tarefa às condições de entrada.

Os resultados indicam que, embora os LLMs ainda não alcancem o nível de concordância observado entre avaliadores humanos, eles apresentam capacidade consistente de aproximação às avaliações humanas, sobretudo em configurações específicas de *prompt*. Além disso, uma con-

tribuição relevante do estudo refere-se à importância da objetividade na formulação e na avaliação das questões. Observou-se que critérios mais claros e bem definidos favorecem não apenas a consistência entre avaliadores humanos, mas também o desempenho dos modelos automáticos, contribuindo para tornar o processo avaliativo mais estruturado. Nesse sentido, a maior objetividade mostrou-se benéfica tanto para o trabalho docente quanto para o desenvolvimento da escrita dos estudantes, além de facilitar o uso de LLMs. Em comparação com abordagens tradicionais ou com o cenário exclusivamente humano, observa-se avanço no sentido de automatizar parcialmente o processo avaliativo, com potencial para aprimorar o diagnóstico e a avaliação de forma integrada entre humanos e máquinas, seguindo a indicação de Rassi & Lopes (2023). Além disso, o estudo analisou um tipo textual (respostas curtas discursivas) menos frequente nos trabalhos de correção automática em português e utilizou dados de estudantes de graduação. Apesar de computacionalmente a amostra ser pequena para considerações genéricas sobre a contribuição de LLMs na tarefa de ASAG, consideramos que linguisticamente é relevante, uma vez que a autorização e a coleta de dados humanos é complexa, exigindo esforço de pesquisa. Reforça-se, com isso, a necessidade de dados anotados de estudantes de diferentes níveis, a fim de proporcionar condições de pesquisa mais eficazes no cenário nacional.

Apesar dos avanços, o trabalho apresenta limitações importantes. Entre elas, destacam-se o tamanho reduzido do conjunto de dados, a restrição a um único tipo de questão discursiva, a ausência de análise de aspectos linguísticos e a ausência de análise qualitativa das justificativas das notas atribuídas pelos modelos — afinal, como se sabe, uma solução completa na tarefa de correção (humana ou automática) precisa vir acompanhada das indicações de melhoria do texto por parte do avaliador. Além disso, a variabilidade no desempenho dos LLMs e a sensibilidade às configurações de *prompt* indicam que os resultados ainda carecem de maior generalização e robustez para aplicação em larga escala.

Como trabalhos futuros, sugere-se a ampliação do *corpus*, contemplando diferentes níveis educacionais, tipos de questões e áreas do conhecimento, bem como a inclusão de análises qualitativas que considerem não apenas a nota, mas também o *feedback* gerado pelos modelos. Adicionalmente, investigações sobre possíveis vieses dos LLMs na atribuição de nota, além de estratégias mais avançadas de *prompt engineering* (como *few-shot* e *chain-of-thought*), e do uso de modelos ajustados especificamente para o português, podem contribuir para reduzir as limitações identificadas. Por fim, a integração de sistemas mais sofisticados, que combinem múltiplos modelos e critérios avaliativos, representa um caminho promissor para o aprimoramento da avaliação automática no contexto educacional.

Agradecimentos

Agradecemos aos estudantes que gentilmente cederam seus textos para análise e aos avaliadores humanos pela atribuição das notas. Também agradecemos aos revisores deste periódico pelas contribuições ao aperfeiçoamento deste artigo. O primeiro autor agradece ao CNPq (processos 420520/2022-8 e 351596/2025-9) e à Fundação Araucária/Seti-PR pelo apoio financeiro a esta pesquisa. Este trabalho foi apoiado pela bolsa Universal do CNPq 2022, pela FAPESC sob o processo 2021TR1510, pela Universidade do Estado de Santa Catarina (UDESC) e indiretamente pelo projeto Céos, financiado pelo Ministério Público do Estado de Santa Catarina (MPSC), que tem contribuído significativamente para o aprimoramento do nosso Laboratório (LISA) e da infraestrutura de processamento de alto desempenho da UFSC. Este trabalho também se enquadra nas atividades do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR).

Contribuições dos autores

Roberlei Alves Bertucci: Conceitualização, Metodologia, Validação, Análise formal, Investigação, Recursos, Redação – rascunho original, Redação – revisão e edição. **Osmar de Oliveira Braz Junior**: Conceitualização, Metodologia, Software, Validação, Análise formal, Investigação, Recursos, Curadoria de dados, Redação – rascunho original, Redação – revisão e edição, Visualização. **Renato Fileto**: Conceitualização, Metodologia, Validação, Análise formal, Investigação, Recursos, Redação – rascunho original, Redação – revisão e edição, Visualização, Supervisão, Administração do projeto.

Referências

Abdullah, Abdulhady Abas, Arkaitz Zubiaga, Seyedali Mirjalili, Amir H Gandomi, Fatemeh Daneshfar, Mohammadsadra Amini, Alan Salam Mohammed & Hadi Veisi. 2025. Evolution of meta's llama models and parameter-efficient fine-tuning of large language models: a survey. arXiv [cs.AI/cs.CL]. 10.48550/arXiv.2510.12178

Abonizio, Hugo, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira & Ramon Pires. 2024. Sabiá-3 technical report. arXiv [cs.CL/cs.AI]. 10.48550/arXiv.2410.12049

Almeida, José Augusto O. da S. & Raimundo Santos Moura. 2024. Investigação de métodos de similaridade textual no contexto da avaliação automática de questões discursivas. Em *Escola Regional de Computação do Ceará, Maranhão e Piauí (ERCEMAPI)*, 110–118.

Bazarim, Milene, Adair Vieira Gonçalves & Gustavo Fechus. 2023. A textualidade em critérios de correção de texto de uma plataforma adaptativa: um estudo de caso. *Revista Letras Raras* 10(2). 124–154. 10.5281/zenodo.10120716

Braz Junior, Osmar de Oliveira, Ana Julia Araujo Sanchuki, Roberlei Alves Bertucci & Renato Fileto. 2025. Verbos dicendi como indicadores de coerência em resumos: uma análise humana e automatizada. *Linguamática* 17(1). 3–16. 10.21814/lm.17.1.461

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell et al. 2020. Language models are few-shot learners. Em *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 1877–1901.

Busato, Lucas Granado. 2025. *Leitura e anotação no ambiente digital: uma pesquisa sobre o desempenho de leitura mediada pela plataforma DLNotes em um contexto pedagógico*: Universidade Tecnológica Federal do Paraná. Tese de Mestrado.

Cançado, Márcia, Luana Amaral, Evelin Amorin, Adriano Veloso & Heliana Mello. 2020. Subjetividade em correções de redações: detecção automática através de léxico de operadores de viés linguístico. *Linguamática* 12(1). 63–79. 10.21814/lm.12.1.313

Cohen, Jacob. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70(4). 213–220. 10.1037/h0026256

Coscarelli, Carla Viana & Luiz Antônio Prazeres. 2013. Avaliando a leitura. Em *Leituras sobre a leitura*, 162–192. Vereda

Dsouza, Nigel. 2025. Comparative analysis of leading generative AI conversational systems: ChatGPT, Grok AI, Gemini, and Meta AI. Authorea Preprints.

Fan, Wenqi, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua & Qing Li. 2024. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. Em *30th ACM Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 6491–6501. 10.1145/3637528.3671470

Ferreira, Paula & Renata Cafardo. 2026. Mais de um terço dos professores avaliados em exame do MEC não sabe o básico para dar aulas. UOL Notícias.

Ferreira Mello, Rafael, Cleon Pereira Junior, Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Newarney Costa, Geber Ramalho & Dragan Gasevic. 2025. Automatic short answer grading in the LLM era: Does GPT-4 with prompt engineering beat traditional models? Em *15th International Learning Analytics and Knowledge Conference (LAK)*, 93–103. 10.1145/3706468.3706481

Galhardi, Lucas, Cinthyan Barbosa, Rodrigo Souza & Jacques Brancher. 2018. Portuguese automatic short answer grading. Em *XXIX Simpósio Brasileiro de Informática na Educação*, 1373–1382. 10.5753/cbie.sbie.2018.1373

Galhardi, Lucas, João Oliveira, Caio Cavazzani & Rodrigo Souza. 2022. AvaliaSmart: Avaliação automática de questões discursivas. Em *XVIII Simpósio Brasileiro de Sistemas de Informação*, 217–220. 10.5753/sbsi_estendido.2022.222233

Galhardi, Lucas, Rodrigo de Souza & Jacques Brancher. 2020. Automatic grading of portuguese short answers using a machine learning approach. Em *XVI Simpósio Brasileiro de Sistemas de Informação*, 109–124. 10.5753/sbsi.2020.13133

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai,

Anja Hauth, Katie Millican et al. 2023. Gemini: a family of highly capable multimodal models. arXiv [cs.CL/cs.AI/cs.CV]. 10.48550/arXiv.2312.11805

Gu, Jindong, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp & Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. arXiv [cs.CV]. 10.48550/arXiv.2307.12980

Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn & Gjergji Kasneci. 2023. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences* 103. 102274. 10.1016/j.lindif.2023.102274

Laitz, Thiago, Thales Sales Almeida, Hugo Abonizio, Roseval Malaquias Junior, Giovana Kerche Bonás, Marcos Piau, Celio Larcher, Ramon Pires & Rodrigo Nogueira. 2026. Sabiá-4 technical report. arXiv [cs.CL/cs.AI]. 10.48550/arXiv.2603.10213

Lalor, John P., Pedro Rodriguez, João Sedoc & Jose Hernandez-Orallo. 2024. Item response theory for natural language processing. Em *18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts (EACL)*, 9–13. 10.18653/v1/2024.eacl-tutorials.2

Leal, Sidney Evaldo, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann & Sandra Maria Aluísio. 2024. NILC-metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation* 58(1). 73–110. 10.1007/s10579-023-09693-w

Lima, Elane do Nascimento Brelaz de. 2025. *Tecnologias da escrita: anotações digitais como estratégia de planejamento textual a partir do uso da plataforma DLNotes*: Universidade Tecnológica Federal do Paraná. Tese de Mestrado.

Liu, Aixin, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong et al. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. arXiv [cs.CL]. 10.48550/arXiv.2512.02556

Marcuschi, Elizabeth. 2004. Avaliação da língua materna: concepções e práticas. *Revista de Letras* 26(1/2). 44–49.

Martins, Ronaldo & Luan Sousa. 2025. Avaliando os avaliadores: quão eficientes são os sistemas de correção automática de redações? *Revista Leia Escola* 25(2). 275–294. 10.35572/rle.v25i2.6493

Meira, Ricardo Radaelli, Augusto Weiand, Eliseo Reategui, Marcio Bigolin & Regina Motz. 2023. A analítica da escrita para identificação de indicadores de qualidade textual. *RENOTE: Revista Novas Tecnologias na Educação* 21(2). 342–351. 10.22456/1679-1916.137756

Mello, Rafael Ferreira, Hilário Oliveira, Moésio Wenceslau, Hyan Batista, Thiago Cordeiro, Ig Ibert Bittencourt & Seiji Isotanif. 2024. PROPOR'24 competition on automatic essay scoring of Portuguese narrative essays. Em *16th International Conference on Computational Processing of Portuguese (PROPOR)*, vol. 2, 1–5.

Monteiro, Edwin & Raimundo Barreto. 2022. Um método baseado na teoria da resposta ao item para avaliação e feedback automático no contexto do ENEM. Em *Simpósio Brasileiro de Informática na Educação (SBIE)*, 255–266. 10.5753/sbie.2022.225182

Mughal, Nimra, Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati & Waheed Noor. 2026. Exploring potential of large language models for automated essay scoring in education. *Discover Artificial Intelligence* 6(1). 166. 10.1007/s44163-026-01002-y

Neha, Fnu & Deepshikha Bhati. 2025. A survey of deepseek models. *Authorea Preprints* 10.36227/techrxiv.173896582.25938392/v1

Nunes, Paula Ávila & Roberlei Alves Bertucci. 2025. Relação entre anotação e desempenho de leitura no ensino superior. *Raído* 19(48). 214–238. 10.30612/raido.v19i48.20263

Oliveira Júnior, Raimundo Cleodimar de. 2025. *ReAvaLIA: um avaliador baseado em modelo de linguagem de grande escala para redações do Enem*: Universidade Federal do Ceará. Tese de Mestrado.

Pearson, Karl. 1895. VII. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58(347–352). 240–242. 10.1098/rspl.1895.0041

Peduzzi, Pedro. 2024. OCDE avalia pensamento criativo de estudantes em 64 países. Agência Brasil, 18 de junho.

Presidência da República. 2025. Enem 2024: resultados mostram crescimento na adesão e na média das notas. Secretaria de Comunicação Social.

Radford, Alec, Karthik Narasimhan, Tim Salimans & Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Relatório técnico. OpenAI.

Rassi, Amanda Pontes & Priscilla de Abreu Lopes. 2023. Correção automática de redação. Em *Processamento de Linguagem Natural Conceitos, Técnicas e Aplicações em Português*, 375–399. Brasileiras em PLN.

Ruiz, Eliana Donaio. 2010. *Como se corrige redação na escola*. Campinas: Mercado de Letras

Santos, Heder Filho Silva. 2025. *Avaliação ayutomática de respostas curtas em Português Brasileiro: Um estudo sobre modelos de linguagem, engenharia de prompt e características textuais*: Instituto Federal Goiano. Tese de Mestrado.

Silva, Alércio Charles. 2023. *Correção automática de questões discursivas de resposta curta: uma abordagem baseada em extração de informação*: Universidade Federal do Maranhão. Tese de Mestrado.

Silveira, Igor Cataneo, Eugénio Ribeiro, Nuno Mamede & Jorge Baptista. 2025. Aprendizado por transferência para correção automática de redação. *Linguamática* 17(2). 99–116. 10.21814/lm.17.2.493

Taschetto, Leonardo de Campos & Renato Fileto. 2025. Evaluating large language models through multidimensional item response theory: A comprehensive case study on ENEM. Em *16th Brazilian Symposium in Information and Human Language Technology (STIL)*, 455–466. 10.5753/STIL.2025.37846

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave & Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models. arXiv [cs.CL]. 10.48550/arXiv.2302.13971

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale et al. 2023b. LLaMA 2: Open foundation and fine-tuned chat models. arXiv [cs.CL/cs.AI]. 10.48550/arXiv.2307.09288

TPE. 2024. Ideb e Saeb: veja os destaques dos resultados de 2023. Todos pela Educação.

UTFPR. 2025. Pesquisa da UTFPR analisa eficácia de anotações no papel e no digital. Universidade Tecnológica Federal do Paraná.

Willmott, Cort J. & Kenji Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 30(1). 79–82. 10.3354/cr030079

Ye, Jiayi, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh Chawla & Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. Em *International Conference on Learning Representations (ICLR)*, 102351–102390. 10.48550/arXiv.2410.02736