Uma Análise Comparativa de LLMs com Técnicas de Engenharia de Prompt para Classificação Automática de Respostas Curtas

Heder Silva Santos[1], Luiz Rodrigues[2], Newarney Torrezão Costa[1], Rafael Ferreira Melo[3], Cleon Pereira Junior[1]

[1]Instituto Federal Goiano

[2]Universidade Tecnológica Federal do Paraná

[3]Universidade Federal Rural de Pernambuco

Heder.filho@estudante.ifgoiano.edu.br

luizrodrigues@utfpr.edu.br

newarney.costa,cleon.junior@ifgoiano.edu.br

rafael.mello@ufrpe.br

Abstract. *Automatic Short Answer Grading (ASAG) aims to reduce human effort in large-scale educational assessments, but there is still limited research in Brazilian Portuguese. This study compares three large language models (GPT-4o-mini, Sabiazinho-3, and Gemini 2.0-Flash) and analyzes the impact of seven prompt engineering elements on their performance. Using a dataset in Brazilian Portuguese, we evaluated all possible combinations of these elements. The combination of few-shot examples with an explicit rubric proved most effective; step-by-step reasoning provided additional benefit specifically for GPT-4o-mini. Sabiazinho-3 showed the highest agreement with human raters, Gemini 2.0-Flash achieved the lowest mean absolute error but produced more hallucinations, and GPT-4o-mini generated cleaner numerical outputs.*

Resumo. *A Correção Automática de Respostas Curtas (ASAG) busca reduzir o esforço humano em avaliações educacionais de larga escala, mas ainda há poucas investigações em português brasileiro. Este estudo compara três grandes modelos de linguagem (GPT-4o-mini, Sabiazinho-3 e Gemini 2.0-Flash) e analisa o impacto de sete elementos de engenharia de prompt no desempenho dos modelos. Com base em um conjunto de dados em português, avaliamos todas as combinações possíveis desses elementos. A combinação de exemplos few-shot com rubrica explícita foi a mais eficaz; o raciocínio passo a passo beneficiou especialmente o GPT-4o-mini. Sabiazinho-3 teve maior concordância com humanos, Gemini 2.0-Flash obteve menor erro médio absoluto, mas com mais alucinações, e o GPT-4o-mini gerou saídas numéricas mais limpas.*

1. Introdução

Avaliar as respostas dos estudantes é uma tarefa fundamental em contextos educacionais, servindo como uma importante fonte de *feedback* que impacta significativamente

DOI: 10.5753/sbie.2025.12705

os resultados de aprendizagem [Burrows et al. 2015]. Tradicionalmente, a avaliação de questões discursivas de resposta curta exige um esforço considerável por parte dos educadores, envolvendo avaliações manuais, resultando em uma atividade dispendiosa [ElNaka et al. 2021]. Esse desafio é particularmente evidente em cenários com inúmeros estudantes, como em Cursos Online Abertos e Massivos (MOOCs), ambientes de ensino a distância e avaliações padronizadas, nos quais a correção rápida e eficaz é crucial, mas é difícil de ser alcançada devido ao alto volume de respostas [Piech et al. 2013].

A Correção Automática de Respostas Curtas (ASAG, do inglês *Automatic Short Answer Grading*) surge como uma solução promissora para mitigar esses problemas, automatizando o processo de avaliação e proporcionando um retorno rápido e consistente aos estudantes [Burrows et al. 2015]. Métodos de ASAG utilizam técnicas computacionais para analisar respostas textuais, verificar sua equivalência semântica em relação às respostas esperadas e atribuir notas automaticamente [Süzen et al. 2020]. Essa abordagem não só reduz significativamente a carga de trabalho dos educadores, como tem o potencial para assegurar consistência e imparcialidade no processo avaliativo.

Os Modelos de Linguagem de Grande Escala (LLMs, do inglês *Large Language Models*) são sistemas de processamento de linguagem natural baseados em redes neurais profundas, treinados em extensos conjuntos de dados [Zhuang et al. 2023]. Modelos anteriores, como o BERT e o Glove, dependiam de aprendizado por transferência e ajuste fino para tarefas específicas, exigindo conhecimento técnico especializado. Diferentemente, os LLMs frequentemente conseguem generalizar melhor entre diferentes tarefas, capturando padrões semânticos e contextuais complexos. Essa capacidade analítica refinada permite que compreendam variações linguísticas, identifiquem nuances e interpretem significados subjacentes nas respostas [Qin et al. 2024]. Porém, pesquisas demonstram que existe um aumento do potencial dessa tecnologia quando associada a ações como ajuste fino e engenharia de *prompts*, que adaptam modelos pré-treinados a domínios ou tarefas específicas [Wei et al. 2021, Carpenter et al. 2024]. Nesse contexto, integrar LLMs à ASAG pode ampliar significativamente a capacidade de compreensão semântica e avaliação contextual dessas respostas.

Estudos recentes em língua inglesa demonstram tanto o potencial quanto os desafios dos LLMs em ASAG. Por exemplo, [Chamieh et al. 2024] investigaram o uso de modelos GPT e LLaMA em cenários *zero-shot* e *few-shot*, comparando-os a modelos supervisionados tradicionais. Os resultados indicaram desempenho insatisfatório das LLMs, especialmente em perguntas que exigiam raciocínio complexo ou conhecimento específico de domínio. Por outro lado, [Grévisse 2024] aplicou a correção automática em um contexto educacional real, onde avaliou o GPT-4 e o Gemini 1.0 corrigindo 2.288 respostas de estudantes de cursos médicos, englobando três idiomas diferentes. Apesar desses avanços, a literatura ainda apresenta lacunas importantes. A predominância de experimentos em língua inglesa, além da pouca investigação sistemática sobre o impacto de diferentes componentes de *prompt*, os custos de implantação e a análise qualitativa de erros são algumas lacunas. Buscando preencher essas lacunas, [Mello et al. 2025] apresentam uma análise comparativa entre modelos tradicionais e o GPT-4, com foco especial em investigar como diferentes estratégias de engenharia de *prompt* podem otimizar o desempenho dos LLMs para a correção automática de respostas curtas em português brasileiro, concluindo que esta abordagem é fundamental para o desempenho do modelo.

Na mesma linha, porém em um estudo focado exclusivamente no português brasileiro, [Mello et al. 2024] avaliaram sistematicamente 128 combinações de *prompt*, reforçando a importância de componentes como *few-shot* e justificativas na eficácia do ASAG em língua portuguesa.

Este trabalho visa investigar especificamente a aplicação de LLMs para ASAG no contexto do português brasileiro. Foram avaliados os modelos: GPT4o-mini, sabiazinho-3 e Gemini 2.0-Flash, analisando suas capacidades de interpretar respostas textuais curtas em português, sendo conduzidos experimentos com técnicas de engenharia de *prompt* para otimizar o desempenho dos modelos, além de fazer uma análise dos componentes de *prompt* usados para a classificação. Nosso objetivo é comparar quais modelos, especificamente abordagens e adaptações, são mais eficazes, contribuindo diretamente para o desenvolvimento de soluções automatizadas mais robustas e precisas no contexto educacional brasileiro.

2. Fundamentação Teórica

Este estudo tem como objetivo analisar diferentes LLMs para Classificação Automática de Respostas Curtas em um conjunto de dados em português brasileiro, considerando também os componentes dos *prompts* utilizados. Nesse sentido, esta seção apresenta uma contextualização geral do ASAG e da engenharia de *prompts*. Em seguida, a seção apresenta trabalhos relacionados sobre aplicações de LLM para ASAG. Finalmente, apresentamos as questões de pesquisa que norteiam este estudo.

2.1. Correção Automática de Respostas Curtas

A Correção Automática de Respostas Curtas (ASAG, do inglês *Automatic Short Answer Grading*) consiste em utilizar métodos de Processamento de Linguagem Natural (PLN) e inteligência artificial para avaliar respostas discursivas breves de estudantes de forma automática [Burrows et al. 2015]. Historicamente, soluções tradicionais de ASAG utilizam técnicas como comparação de similaridade textual, análise de *bag-of-words* ou técnicas de PLN para extrair termos-chave e padrões de linguagem [Ripmiatin et al. 2024].

Ao longo das últimas décadas, diversas abordagens técnicas foram desenvolvidas para viabilizar o ASAG. Métodos baseados em similaridade textual e semântica foram pioneiros. Por exemplo, [Mohler and Mihalcea 2009] exploraram medidas de similaridade semântica (usando recursos como WordNet e LSA) para comparar a resposta do aluno com a resposta de referência, alcançando correlação de 0,50 entre o sistema e o humano, em comparação a 0,64 de concordância entre dois humanos.

Com o avanço do aprendizado profundo, o campo de ASAG passou por uma evolução significativa. Modelos de word embeddings e redes neurais passaram a ser empregados para capturar melhor o contexto e o significado das respostas de estudantes [Ahmed et al. 2022]. Diversos estudos exploraram arquiteturas de redes neurais para ASAG. [Camus and Filighera 2020] investigaram o uso de transformadores utilizando o modelo BERT para pontuação automática, enquanto [Sung et al. 2019] reportaram melhorias no desempenho de tarefas de ASAG ao pré-treinar modelos do tipo *transformer* em dados educacionais.

Além dos desafios que a própria tarefa de ASAG apresenta, outro ponto notório é a escassez de pesquisas sobre sua aplicação e teste no contexto do português brasi-

leiro [Galhardi et al. 2020]. A grande parte dos estudos e do desenvolvimento na área de ASAG concentra-se no idioma inglês [Burrows et al. 2015]. Como resultado, há menos conjuntos de dados prontos e disponíveis em português, menos modelos de linguagem treinados para os detalhes específicos da educação no Brasil e poucos estudos que tratem das características próprias da língua [Galhardi et al. 2018].

2.2. Modelos de Linguagem de Grande Escala

Os LLMs representam um avanço significativo no campo do Processamento de Linguagem Natural (PLN), pois são baseados em arquiteturas de redes neurais profundas, como o transformer, que empregam mecanismos de atenção para modelar dependências de longo alcance em textos [Vaswani et al. 2023]. Diferentemente das abordagens precedentes, que exigiam pré-processamento extensivo e engenharia manual de características, os LLMs são treinados em regimes auto-supervisionados com grandes volumes de dados, desenvolvendo assim representações contextuais robustas da linguagem [Zhao et al. 2025].

Para a ASAG, os LLMs oferecem oportunidades potenciais e significativas sobre as abordagens anteriores [Mello et al. 2025]. Sua profunda compreensão semântica permite avaliar respostas com base no significado, em vez de somente na forma lexical, tornando-os mais aptos a reconhecer respostas corretas expressas de maneiras inesperadas [Yan et al. 2024]. Além disso, muitos LLMs demonstram fortes capacidades de aprendizado em poucos exemplos (*few-shot learning*) ou mesmo sem exemplos (*zero-shot learning*), podendo adaptar-se a novas tarefas de avaliação com instruções em linguagem natural, sem a necessidade de grandes conjuntos de dados de treinamento específicos para cada questão. A capacidade de seguir instruções complexas e gerar não somente notas, mas também justificativas ou *feedback*.

Mais recentemente, a emergente era dos LLMs influencia o panorama do ASAG. LLMs como o GPT-3 e GPT-4, treinados em corpora massivos e capazes de compreensão contextual avançada, abriram caminho para abordagens baseadas em engenharia de *prompt* ao invés de treinamento supervisionado tradicional [Mello et al. 2025, Mello et al. 2024]. Por exemplo, estudos demonstraram que os modelos GPT-3.5 e GPT-4 podem ser utilizados para avaliar respostas em finlandês com desempenho competitivo com métodos anteriores, mesmo sem treinamento adicional específico naquela língua [Chang and Ginter 2024].

Apesar do potencial transformador dos LLMs para tarefas de ASAG, sua adoção ainda enfrenta limitações importantes. Diversos estudos evidenciam que, embora capazes de alcançar desempenho competitivo, esses modelos podem apresentar inconsistências, enviesamentos e uma propensão à geração de respostas alucinatórias — ou seja, conteúdos plausíveis, mas incorretos ou irrelevantes para a tarefa [Xu et al. 2025, Bang et al. 2023]. Essa tendência é especialmente preocupante em cenários educacionais, onde avaliações automáticas exigem confiabilidade e transparência. Além disso, pesquisas evidenciam que LLMs podem ser sensíveis à formulação dos *prompts* e à estrutura das instruções fornecidas, impactando diretamente a qualidade e a precisão das respostas geradas [Zhao et al. 2025, Mello et al. 2024].

2.3. Engenharia de prompt

A engenharia de *prompt* surgiu como uma disciplina essencial no uso de LLMs, especialmente com o avanço das diversas arquiteturas [Khot et al. 2023]. Ela consiste na

elaboração estratégica de entradas textuais visando induzir comportamentos, saídas e formatações específicas no modelo [Liu et al. 2023]. Essa prática se tornou especialmente relevante ao se observar que pequenas mudanças na formulação de um *prompt* podem impactar significativamente a qualidade, precisão e utilidade das respostas geradas [Karmaker Santu and Feng 2023].

O propósito da engenharia de *prompt* é, portanto, maximizar o desempenho dos modelos sem a necessidade de ajustes nos parâmetros internos. Isso é alcançado por meio de diversas técnicas que buscam aprimorar o desempenho de LLMs [Karmaker Santu and Feng 2023]. Entre elas estão *few-shot prompting*, que apresenta exemplos de entrada–saída como referência; instruções detalhadas, listando claramente objetivos e critérios [Carpenter et al. 2024]; *chain-of-thought*, que exige raciocínio em etapas antes da resposta final [Wei et al. 2022]; definição de persona, atribuindo ao modelo um papel específico para guiar o tom e o rigor; e decomposição modular de tarefas, que divide problemas complexos em sub-tarefas resolvidas sequencialmente [Khot et al. 2023].

Pesquisas recentes em ASAG confirmam o peso da engenharia de *prompt*. [Mello et al. 2024] avaliaram 128 variações de *prompt* em português com GPT-3.5 e GPT-4 e descobriram que inserir um "tempo para pensar" e exigir justificativa da nota elevou sistematicamente o desempenho, com o GPT-4 superando o GPT-3.5 quando guiado por *prompts* bem estruturados. Em estudo semelhante, [Chang and Ginter 2024] mostraram que, no ChatGPT, a inclusão de um exemplo esperado aliado a instruções claras de formato aumentou a concordância com avaliadores humanos em finlandês, superando a configuração *zero-shot*. Em conjunto, esses achados evidenciam que componentes como exemplos, instruções precisas e raciocínio explícito

Vale notar, por fim, considerações específicas sobre a língua portuguesa na engenharia de *prompt*. Como muitos LLMs foram treinados predominantemente em inglês, a efetividade dos *prompts* em português depende não só das técnicas mencionadas, mas também de ajustes linguísticos [Freitag and Gois 2024]. Pesquisas recentes supriram a falta de estudos focados nesse contexto, onde forneceram diretrizes valiosas para elaborar *prompts* eficazes em português brasileiro, demonstrando na prática quais componentes do *prompt* mais contribuem para melhorar a acurácia da correção automática no idioma [Mello et al. 2025]

2.4. Questões de Pesquisa

Mesmo com avanços consideráveis na última década, a literatura sobre ASAG ainda apresenta lacunas importantes, sobretudo quando se trata de (i) compreender, de forma sistemática, quais elementos de *prompt* mais impactam o desempenho de LLMs na tarefa e (ii) comparar o desempenho entre múltiplos LLMs quando avaliados em respostas em português brasileiro. Pesquisas anteriores em português brasileiro têm se limitado, em geral, a um único modelo de referência, como o GPT-4, deixando em aberto o quanto as arquiteturas distintas, incluindo modelos nativos para a língua, podem contribuir para a tarefa de ASAG. Com esses desafios em vista, formulamos as seguintes questões de pesquisa:

Questão de Pesquisa 1 (QP1): *Qual o desempenho de diferentes modelos de LLMs no contexto de ASAG para português brasileiro?*

Questão de Pesquisa 2 (QP2): *Quão suscetíveis esses modelos são à geração de alucinações durante a tarefa de ASAG?*

Questão de Pesquisa 3 (QP3): *Quais componentes específicos do design de prompt podem aumentar a efetividade de LLMs quando aplicados a ASAG?*

Embora pesquisas recentes já tenham comparado GPT-3.5, GPT-4 e modelos open-source em línguas distintas [Chang and Ginter 2024], faltam investigações focadas em português brasileiro. Por sua vez, investigações em português brasileiro são limitadas aos modelos GPT [Mello et al. 2024, Mello et al. 2025]. As QPs 1 e 2, portanto, pretendem mensurar a qualidade dos modelos com arquiteturas e tamanhos diferentes quando expostos à tarefa de atribuir notas objetivas a respostas dissertativas curtas. Assim, assas questões são complementares: QP3 foca na engenharia de *prompt*, enquanto QP2 e QP1 examina a variação inerente às próprias arquiteturas de LLM.

3. Metodologia

3.1. Dataset

O conjunto de dados utilizado neste estudo, denominado PT_ASAG, foi proposto por [Galhardi et al. 2020]. Ele é composto por 7.473 respostas textuais curtas fornecidas por 659 estudantes em resposta a 15 questões de Biologia, todas formuladas em português brasileiro. Neste conjunto, 14 estudantes de graduação em Biologia avaliaram as respostas utilizando uma escala predefinida. Cada resposta foi avaliada por pelo menos dois estudantes, alcançando um índice de concordância entre os avaliadores de 0,43 segundo a estatística kappa de Cohen.

Assim como em [Mello et al. 2024], aplicamos aproximadamente 30% do volume de dados original, correspondendo a cerca de 2.641 respostas (Tabela 1). Essa abordagem teve como objetivo otimizar a viabilidade computacional para análises exploratórias e testes intensivos de engenharia de *prompts*, ao mesmo tempo que assegurou representatividade suficiente para inferir sobre o desempenho dos modelos na tarefa proposta.

Tabela 1. Estatísticas do conjunto de dados utilizado

	Dados totais	Dados usados
Questões	15	15
Respostas	7.473	2.641

3.2. Modelos de LLM

Para investigar o desempenho da Correção Automática de Respostas Curtas em português, selecionamos três LLMs cujos perfis se complementam tanto em arquitetura quanto em contexto de uso:

- **GPT.4o-mini-2024-07-18** (OpenAI)[1], escolhido por já possuir estudos prévios aplicados a ASAG, o que viabiliza comparações diretas com a literatura. A variante mini mantém o núcleo de raciocínio avançado do GPT-4o completo, incluindo atenção a múltiplas modalidades e janela de contexto estendida, porém a um custo computacional mais baixo, fator importante para experimentos repetidos [Brown et al. 2020].

[1] https://platform.openai.com/

- **Sabiazinho-3** (Maritaca AI)[2], representa a vertente nacional, por ser um modelo treinado e afinado em bases predominantemente em português brasileiro. Tal especialização tende a captar nuances linguísticas e culturais que impactam a atribuição de notas em respostas escritas em português, além de oferecer menor custo por *token* [Abonizio et al. 2025b].
- **Gemini 2.0-Flash** (Google DeepMind)[3], incluído como contraponto leve (*flash*) para escalonar testes em larga escala. Embora tenha menos recursos multimodais, ele combina janela de contexto ampla com inferência rápida e econômica, facilitando a execução de centenas de *prompts* em paralelo [Imran and Almusharraf 2024].

Todas as requisições foram realizadas via API, mantendo o parâmetro *temperature* = 0,0, com exceção do sabiazinho-3 que foram mantidos os parâmetros como recomendados na documentação. Com isso, foi possível eliminar incertezas e focar na capacidade de avaliação objetiva. O número máximo de tokens foi restringido ao intervalo de 5, pois o objetivo era coletar exclusivamente a nota atribuída, sem texto explicativo. Os valores dos parâmetros foram obtidos via testes empíricos, conforme a Tabela 2.

Tabela 2. Parâmetros de inferência empregados em cada LLM

Modelo	Temperatura	Max. Tokens
GPT-4o-mini	0,0	5
Sabiazinho-3	0,9	5
Gemini 2.0-Flash	0,0	5

3.3. Engenharia de Prompt

Para investigar de forma sistemática o efeito da engenharia de *prompt* no desempenho dos LLMs em tarefas de correção automática de respostas curtas, adotamos uma abordagem baseada em composição-decomposição modular dos *prompts* [Mello et al. 2025]. Aderimos também a medidas e práticas recomendadas na literatura [White et al. 2023]. Primeiramente, definimos dois elementos fixos presentes em todas as instruções: (i) a Instrução, que descreve a escala de avaliação e o objetivo da tarefa, e (ii) o Formato de Saída, que determina que o modelo devolva somente a nota final [Giray 2023]. Outras estratégias incluem permitir que o modelo "pense" antes de responder [Khot et al. 2023]; atribuir ao modelo um papel ou persona específica para orientar sua geração de texto; apresentar exemplos de interações corretas (*few-shot*), fornecer informações adicionais ou contexto suplementar [Wei et al. 2022], entre outras abordagens.

Visando maximizar a acurácia dos LLMs na tarefa de correção, foi conduzida uma investigação sistemática sobre a formulação do *prompt* de instrução. Uma estrutura de *prompt* base foi decomposta em sete componentes modulares independentes, com dois componentes sendo obrigatórios, sendo eles a Instrução e a saída. Cada componente representa um elemento potencial da instrução, conforme a tabela 3.

Para determinar a configuração de *prompt* mais eficaz para cada modelo, foram geradas e testadas todas as 128 (2^7) combinações possíveis, resultantes da inclusão ou

[2]https://plataforma.maritaca.ai/
[3]https://ai.google.dev/

Tabela 3. Descrição dos componentes de *prompt*

Componentes	Texto em português
instrução	Avalie a resposta dos alunos numa escala de 0 (completamente errado) a 3 (resposta perfeita).
contexto	Você esta corrigindo uma atividade do ensino médio.
papel	Assuma o papel de um professor de ensino médio.
tempo para pensar	Pense passo a passo.
passo a passo	Siga os seguintes passos: 1. formule a sua resposta para a pergunta. 2. verifique se todos os itens relevantes que você identificou estão na resposta do aluno. 3. elabore um racional para justificar a qualidade da resposta do aluno. 4. Compare a sua resposta e o seu racional com a do aluno para dar a nota final.
few-shot	Utilize o exemplo abaixo de resposta correta na sua correção Questão: `question_instructor` Resposta: `answer_instructor`.
rubrica	A nota final deve avaliar se o conteúdo foi respondido na sua correção.
justificativa	Raciocine sobre a justificativa para sua avaliação explicando suas decisões para a nota final.
saída	O resultado deve ser apenas a nota final: 0, 1, 2 ou 3.

exclusão de cada um dos sete componentes. Cada combinação formou um *prompt* final distinto, que foi então utilizado para instruir os LLMs a avaliar as respostas curtas do conjunto de dados.

3.4. Avaliação

A validação dos modelos foi estruturada em três métricas complementares, o Coeficiente de Concordância de Cohen (k), Erro Médio Absoluto (MAE) e Erro Quadrático Médio (RMSE) de modo a analisar tanto a consistência qualitativa quanto a precisão quantitativa dos resultados. O Coeficiente k quantifica o acordo entre as notas geradas pelos modelos e aquelas atribuídas por avaliadores humanos, ajustando-se pela concordância aleatória e, assim, fornecendo uma medida robusta de confiabilidade interavaliador; valores mais elevados de κ sinalizam maior consistência do modelo em replicar o julgamento humano [Fleiss et al. 1969]. Por sua vez, o MAE reflete a discrepância média absoluta entre predição e referência, oferecendo uma interpretação direta do erro médio sem penalizar desproporcionalmente desvios extremos [Zhao et al. 2017]. Por fim, o RMSE acentua os maiores erros ao elevar ao quadrado as diferenças individuais, revelando até que ponto erros atípicos podem comprometer a robustez do resultado [Zhao et al. 2017].

Para responder às questões de pesquisa 1 e 3, aplicaram-se as três métricas de maneira independente a cada modelo e configuração de *prompt* remanescente, permitindo comparar não apenas o desempenho absoluto dos modelos, mas também aprofundar a análise de impactos individuais de cada elemento de *prompting*. Assim, foi elaborado um ranking dos componentes dos *prompts*. Nesse processo, cada combinação dos componentes foi avaliada isoladamente pelo nível de κ, MAE e RMSE. A partir dessas combinações,

definiram-se três faixas de relevância: Top-5, Top-10 e Top-20. Com isso, foi possível identificar tendências entre os componentes mais prevalentes nos elementos de *prompts* de maior pontuação.

Para responder a questão de pesquisa 2, foi avaliado quantas vezes as arquiteturas de LLM apresentaram a geração de texto não-numérico em resposta aos *prompts* que deveriam produzir somente uma nota [Rawte et al. 2023, Xu et al. 2025], caracterizando as alucinações neste caso. Para assegurar a relevância estatística das métricas, estabeleceu-se como critério de condição que cada configuração de *prompt* admitisse no máximo 25% de alucinações. A adoção do corte em 25% mantém o poder estatístico suficiente, preservando pelo menos três quartos das respostas em cada condição experimental para cálculos confiáveis do κ, MAE e RMSE [Warneke et al. 2025]. Dessa forma, equilibra-se a necessidade de representatividade e a integridade dos resultados, impedindo que falhas de interpretação isoladas comprometam as conclusões sobre a eficácia das diferentes estratégias de engenharia de *prompt*.

4. Resultados

4.1. QP1: Em que medidas e como desempenham os distintos modelos de LLMs no contexto de ASAG?

Entre os *prompts* em que os modelos alcançaram maior concordância categórica com o avaliador humano, o sabiazinho-3 destacou-se. Alcançando um κ médio de 0,50, variando de 0,49 a 0,50, sabiazinho-3 foi superior aos demais. Esse valor indica que, sob condições ideais de *prompt*, ele é o modelo que mais se alinha às decisões humanas. Embora não lidere no MAE e RMSE o sabiazinho-3, ainda possui a melhor média de MAE 0,49, variando de 0,38 a 0,40, sugerindo que, além de classificar corretamente a maioria das categorias, quando erra, o desvio tende a ser menos acentuado.

Por sua vez, o Gemini 2.0-Flash aproxima-se do desempenho do sabiazinho-3, alcançando κ = 0,49 com somente "contexto + rubrica" ou "contexto + papel + rubrica", porém com ligeiro ganho em precisão numérica (MAE = 0,37; RMSE = 0,68) frente ao modelo em português brasileiro. Por fim, o GPT-4o-mini figura agora na terceira posição, com κ variando de 0,37 a 0,41. Seus melhores resultados surgem quando combinamos raciocínio guiado ("tempo para pensar + passo a passo") a exemplos *few-shot* e rubrica.

4.2. QP2: Quão suscetíveis esses mesmos modelos são à geração de alucinações durante a tarefa?

A Tabela 5 revela um quadro do nível de alucinação de cada LLM na tarefa de ASAG. O GPT-4o-mini se destaca por quase não apresentar alucinações, produzindo em média 0,76% de linhas alucinatórias por configuração de *prompt*. Assim, nenhuma configuração precisou ser descartada. O sabiazinho-3, embora apresente uma média de alucinação maior por *prompt* (8,30%), ainda opera confortavelmente abaixo do limite de 25% adotado como corte neste estudo. Já o Gemini 2.0-Flash destaca-se negativamente, alucinado em média quase metade das linhas (48,5%) operando muito acima do limite estipulado, assim, exigindo a exclusão de 76 linhas que superaram este limite. Entre as linhas descartadas, observamos trechos como "Vamos analisar...", "Analisando..." ou "Para aval...", saídas discursivas que fogem inteiramente do formato numérico esperado e ilustram como respostas verbosas que contaminam a métrica. Esses resultados sugerem que, ao menos

Tabela 4. Desempenho de diferentes combinações de componentes de *prompt*

Componentes do *Prompt*	Modelo	k	MAE	RMSE
few-shot + justificativa	sabiazinho-3	0,50	0,38	0,71
few-shot + rubrica + justificativa	sabiazinho-3	0,50	0,38	0,71
contexto + *few-shot* + rubrica	sabiazinho-3	0,49	0,40	0,75
papel + tempo para pensar + *few-shot* + rubrica + justificativa	sabiazinho-3	0,49	0,39	0,71
contexto + papel + *few-shot* + rubrica	sabiazinho-3	0,49	0,39	0,72
tempo para pensar + passo a passo + *few-shot* + rubrica	GPT4o	0,41	0,50	0,86
tempo para pensar + passo a passo + *few-shot* + rubrica + justificativa	GPT4o	0,39	0,49	0,86
contexto + tempo para pensar + passo a passo + *few-shot* + rubrica	GPT4o	0,38	0,54	0,93
tempo para pensar + passo a passo + *few-shot* + justificativa	GPT4o	0,37	0,51	0,89
papel + tempo para pensar + passo a passo + *few-shot* + rubrica	GPT4o	0,37	0,57	0,98
contexto + papel + rubrica	Gemini2.0-Flash	0,49	0,37	0,68
contexto + rubrica	Gemini2.0-Flash	0,49	0,37	0,68
rubrica	Gemini2.0-Flash	0,48	0,41	0,73
contexto + passo a passo + *few-shot* + rubrica	Gemini2.0-Flash	0,46	0,45	0,83
papel + rubrica	Gemini2.0-Flash	0,46	0,41	0,72

para este experimento, versões "*flash*" que priorizam latência e economia de tokens podem sacrificar a fidelidade das respostas.

Tabela 5. Média de alucinações (MA) e quantidade de configurações de *prompt* (QC) removidas dos experimentos para cada modelo.

Modelo	MA	QC
GPT4o-mini	0,76%	0
sabiazinho-3	8.30%	17
Gemini 2.0-Flash	48.50%	76

4.3. QP3: Quais componentes específicos do design de *prompt* podem aumentar a efetividade de LLMs quando aplicados a ASAG?

A Tabela 6 contabiliza quantas vezes cada componente aparece entre os Top-5, Top-10 e Top-20 *prompts* de maior κ para cada modelo. Destaca-se que *few-shot* está presente em quase todas as ocorrências dos três modelos configurações que explicitam os critérios de avaliação por meio da rubrica encontram-se de maneira predominante entre os Top-10 de cada modelo. O padrão passo a passo destaca-se no GPT-4o-mini (13 ocorrências no Top-20) mas é marginal no Gemini 2.0-Flash, enquanto não aparece no sabiazinho-3. Em contraste, o componente papel, concebido para situar o modelo em uma determinada persona, manifesta impacto marginal na maior parte dos cenários.

Tabela 6. Contagem de ocorrência de componentes nos Top-5 (T5), Top-10 (T10) e Top-20 (T20) para cada modelo.

Componente	GPT4o-mini			Gemini2.0-flash			sabiazinho-3		
	T5	**T10**	**T20**	**T5**	**T10**	**T20**	**T5**	**T10**	**T20**
contexto	1	4	9	3	5	10	2	4	12
few-shot	**5**	**10**	**19**	1	3	13	**5**	**10**	**20**
passo a passo	**5**	6	13	1	1	3	0	0	0
rubrica	4	9	18	**5**	**8**	**14**	4	9	14
justificativa	3	3	7	0	0	2	3	5	9
papel	1	2	7	2	4	5	2	5	9
tempo para pensar	**5**	8	14	0	1	5	1	2	8

5. Discussão

Os resultados obtidos neste estudo elucidam dois eixos centrais: (i) a relevância de como se formula o *prompt* para tarefas de ASAG e (ii) o comportamento de diferentes arquiteturas de LLM diante de métricas distintas de avaliação.

Os achados deste estudo destacam, a importância crítica da engenharia de *prompt* para o desempenho de LLMs em tarefas de ASAG. A análise da frequência dos componentes nos *prompts* mais eficazes (Tabela 6) revela que a combinação de *few-shot* e rubrica se estabelece como fator determinante para maximizar a concordância com avaliadores humanos. Esses elementos fornecem, respectivamente, exemplos concretos de respostas corretamente avaliadas e um guia normativo explícito dos critérios de correção. O *few-shot* circunscreve o espaço de possíveis rótulos, e ao explicitar os requisitos de cada categoria de nota, a rubrica ancora o raciocínio do LLM em expectativas objetivas, limitando interpretações subjetivas [Carpenter et al. 2024]. Em contraste, o componente papel mostrou impacto marginal em quase todas as configurações, sugerindo que personificações isoladas não são suficientes para melhorar a qualidade de correção quando não apoiadas por instruções e critérios claros.

Do ponto de vista comparativo entre modelos (QP1), o sabiazinho-3 exibiu o maior κ médio de 0,49 (máximo de 0,50) e em média a menor dispersão de notas, indicando elevada consistência categórica e baixa propensão a erros extremos. Esse comportamento ratifica a hipótese de que o modelo ser treinado para português brasileiro confere vantagens decisivas em tarefas ligadas a texto acadêmico na língua portuguesa, ainda que o modelo opere com menos parâmetros do que seus concorrentes globais [Abonizio et al. 2025a]. O Gemini 2.0-Flash, por sua vez, atingiu os melhores valores médios de MAE, aproximadamente 0,37, sinalizando boa precisão absoluta, mas oscilou mais em κ. Já o GPT-4o-mini manteve desempenho intermediário em ambas as métricas. Por outro lado, o destaque ao GPT-4o reside em sua robustez contra alucinações comparado ao sabiazinho-3 e, principalmente, Gemini.

Nesse contexto, este estudo converge com a literatura ao expandir trabalhos anteriores, reafirmando que a combinação de exemplos *few-shot* e instruções explícitas de rubrica constitui o núcleo das estratégias de engenharia de prompt mais eficazes para ASAG [Mello et al. 2024]. Nossa investigação demonstra que essa abordagem dupla maximiza a concordância com avaliadores humanos. Nossos melhores resultados, obtidos com o

modelo sabiazinho-3, apresentaram um coeficiente κ de 0,49 valor compatível com os resultados da literatura no conjunto de dados PT_ASAG [Galhardi et al. 2020], que também reporta um κ médio de 0,49. Por outro lado, nossos melhores desempenhos em MAE, alcançados com o sabiazinho-3 e Gemini 2.0-Flash (MAE médio de 0,38 e 0,40, respectivamente), superam resultados de trabalhos recentes que utilizam modelos clássicos, como variantes de TF-IDF, que obtêm um MAE médio de 0,42 [Mello et al. 2025]. Além disso, os resultados indicam que variações de *prompting* como passo a passo ou *time-to-think* apresentam impacto dependente da arquitetura do modelo e do idioma [Mello et al. 2024].

Com isso, este artigo se distingue dos estudos antecessores por frentes complementares. Primeiro, ele amplia o escopo de análise ao comparar diretamente três arquiteturas de LLM (GPT-4o-mini, Sabiazinho-3 e Gemini 2.0-Flash), enquanto investigações prévias em português brasileiro focavam quase exclusivamente em variantes do GPT. Porém, ainda há necessidade de pesquisas em torno das LLMs para que possam ser aplicadas de forma eficaz na classificação automática de respostas curtas.

6. Limitações e trabalhos futuros

Embora os resultados desta investigação ofereçam evidências promissoras sobre o uso de LLMs para ASAG em português brasileiro, é importante reconhecer alguns fatores que restringem o seu escopo. Em primeiro lugar, a análise utilizou aproximadamente 30% do conjunto original para tornar os testes de engenharia de *prompts* viáveis em tempo e financeiramente. Essa amostragem, ainda que supere em número alguns conjuntos da literatura, é limitada ao domínio de Biologia de nível médio, deixando em aberto o desempenho em outras disciplinas, faixas etárias ou estilos de escrita.

Apesar dos nossos resultados positivos, observamos taxas de alucinação superiores a 25% em algumas combinações de *prompts* um fenômeno amplamente documentado como limitação intrínseca dos LLMs [Xu et al. 2025]. Nosso critério de corte mitigou o impacto desses casos sobre as métricas, mas reforça a necessidade de validação automatizada da saída antes de adoção em cenários de alto risco educativo. Trabalhos futuros deveriam explorar melhores parâmetros para a chamada da API, bem como técnicas de pós-processamento de *prompts*, validação automática de saída e abordagens híbridas, visando reduzir ainda mais a incidência de alucinações e melhorar a robustez geral dos sistemas de ASAG em português brasileiro.

Outro ponto limitador reside no caráter estático dos *prompts* utilizados. Embora tenhamos testado 128 combinações, todas assumem um cenário de correção em que o modelo recebe a pergunta e devolve a nota sem interação posterior. Isso contrasta com situações autênticas de sala de aula, nas quais o professor pode solicitar justificativas, refutações ou sugestões de melhoria. Investigações futuras poderiam explorar esquemas *multi-turn*, nos quais o LLM revisa sua nota após contra-argumentos do aluno, além de comparar estratégias puramente de engenharia de *prompt* com abordagens de aprendizado supervisionado ou de geração aumentada de recuperação. Essas linhas de pesquisa permitiriam aferir se a combinação de *feedback* iterativo e adaptação de domínio reduz a incidência de alucinações, ao mesmo tempo, em que melhora a transparência e a utilidade pedagógica.

7. Disponibilidade de Artefatos

Os artefatos deste estudo são disponibilizados através de contato aos autores correspondentes.

Referências

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2025a). Sabiá-3 technical report.

Abonizio, H. et al. (2025b). Sabiá-3 technical report. Preprint.

Ahmed, A., Joorabchi, A., and Hayes, M. J. (2022). On deep learning approaches to automated assessment: Strategies for short answer grading. *CSEDU (2)*, pages 85–94.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Burrows, S., Gurevych, I., and Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25:60–117.

Camus, L. and Filighera, A. (2020). Investigating transformers for automatic short answer grading. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II*, page 43–48, Berlin, Heidelberg. Springer-Verlag.

Carpenter, D., Min, W., Lee, S., Ozogul, G., Zheng, X., and Lester, J. (2024). Assessing student explanations with large language models using fine-tuning and few-shot learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 403–413, Mexico City, Mexico. Association for Computational Linguistics.

Chamieh, I., Zesch, T., and Giebermann, K. (2024). LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z., editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 309–315, Mexico City, Mexico. Association for Computational Linguistics.

Chang, L.-H. and Ginter, F. (2024). Automatic short answer grading for finnish with chatgpt. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23173–23181.

ElNaka, A., Nael, O., Afifi, H., and Sharaf, N. (2021). Arascore: Investigating response-based arabic short answer scoring. *Procedia Computer Science*, 189:282–291. AI in Computational Linguistics.

Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327.

Freitag, R. M. K. and Gois, T. S. d. (2024). Performance in a dialectal profiling task of llms for varieties of brazilian portuguese. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2024)*, STIL 2024, page 317–326. Sociedade Brasileira de Computação.

Galhardi, L., Barbosa, C., Thom de Souza, R. C., and Brancher, J. (2018). Portuguese automatic short answer grading. page 1373.

Galhardi, L., de Souza, R., and Brancher, J. (2020). Automatic grading of portuguese short answers using a machine learning approach. In *Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação*, pages 109–124.

Giray, L. (2023). Prompt engineering with chatgpt: A guide for academic writers. *Annals of Biomedical Engineering*, 51(12):2629–2633.

Grévisse, C. (2024). Llm-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, 24(1):1060.

Imran, M. and Almusharraf, N. (2024). Google gemini as a next generation ai educational tool: A review of emerging educational technology. *Smart Learning Environments*, 11(1):22.

Karmaker Santu, S. K. and Feng, D. (2023). TELeR: A general taxonomy of LLM prompts for benchmarking complex tasks. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14197–14203, Singapore. Association for Computational Linguistics.

Khot, T. et al. (2023). Decomposed prompting: A modular approach for solving complex tasks. Preprint.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Mello, R., Rodrigues, L., Cabral, L., Pereira, F., Júnior, C. P., Gasevic, D., and Ramalho, G. (2024). Prompt engineering for automatic short answer grading in brazilian portuguese. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 1730–1743, Porto Alegre, RS, Brasil. SBC.

Mello, R. F., Pereira Junior, C., Rodrigues, L., Pereira, F. D., Cabral, L., Costa, N., Ramalho, G., and Gasevic, D. (2025). Automatic short answer grading in the llm era: Does gpt-4 with prompt engineering beat traditional models? In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*, pages 93–103, New York, NY, USA. Association for Computing Machinery.

Mohler, M. and Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In Lascarides, A., Gardent, C., and Nivre, J., editors, *Proceedings of*

the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 567–575, Athens, Greece. Association for Computational Linguistics.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. (2013). Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*.

Qin, L., Chen, Q., Feng, X., Wu, Y., Zhang, Y., Li, Y., Li, M., Che, W., and Yu, P. S. (2024). Large language models meet nlp: A survey. Preprint.

Rawte, V., Sheth, A., and Das, A. (2023). A survey of hallucination in large foundation models.

Ripmiatin, E., Purnamasari, P. D., and Ratna, A. A. P. (2024). Comparing classical distance measures and word embeddings for automatic short answer grading. In *Proceedings of the 2023 9th International Conference on Communication and Information Processing*, ICCIP '23, page 492–497, New York, NY, USA. Association for Computing Machinery.

Sung, C., Dhamecha, T. I., and Mukhi, N. (2019). Improving short answer grading using transformer-based pre-training. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I*, page 469–481, Berlin, Heidelberg. Springer-Verlag.

Süzen, N., Gorban, A. N., Levesley, J., and Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. In *Procedia Computer Science*, volume 169, pages 726–743.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

Warneke, K., Keiner, M., Wallot, S., Siegel, S. D., Günther, C., Wirth, K., and and, S. P.-M. (2025). The impact of sample size on reliability metrics stability in isokinetic strength assessments: Does size matter? *Measurement in Physical Education and Exercise Science*, 0(0):1–12.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Xu, Z., Jain, S., and Kankanhalli, M. (2025). Hallucination is inevitable: An innate limitation of large language models.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., and Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55:90–112.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2025). A survey of large language models.

Zhao, Z., Lang, W., Doulgeris, A. P., and Chen, L. (2017). Improved llm methods using linear regression. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5350–5353, Fort Worth, TX, USA. IEEE.

Zhuang, Z., Chen, Q., Ma, L., Li, M., Han, Y., Qian, Y., Bai, H., Feng, Z., Zhang, W., and Liu, T. (2023). Through the lens of core competency: Survey on evaluation of large language models.