BACHAREL EM CIÊNCIA DA COMPUTAÇÃO

AVALIAÇÃO AUTOMÁTICA DE RESPOSTAS CURTAS EM PORTUGUÊS BRASILEIRO: UM ESTUDO SOBRE MODELOS DE LINGUAGEM, ENGENHARIA DE PROMPT E CARACTERÍSTICAS TEXTUAIS

HEDER FILHO SILVA SANTOS

Iporá, GO

2025

INSTITUTO FEDERAL GOIANO - CAMPUS IPORÁ
BACHAREL EM CIÊNCIA DA COMPUTAÇÃO

AVALIAÇÃO AUTOMÁTICA DE RESPOSTAS CURTAS EM PORTUGUÊS BRASILEIRO: UM ESTUDO SOBRE MODELOS DE LINGUAGEM, ENGENHARIA DE PROMPT E CARACTERÍSTICAS TEXTUAIS

HEDER FILHO SILVA SANTOS

Trabalho de Conclusão de Curso apresentado ao Instituto Federal Goiano - Campus Iporá, como requisito parcial para a obtenção do Grau de Bacharel em Ciência da Computação.

Orientador: Prof. Prof. Dr. Cleon Xavier Pereira Junior
Coorientador: Prof. Dr. Luiz Antonio Lima Rodrigues
Universidade Tecnológica Federal do Paraná

Iporá, GO

dezembro, 2025

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO

PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610, de 19 de fevereiro de 1998, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano a disponibilizar gratuitamente o documento em formato digital no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

- ☐ Tese (doutorado)
- ☐ Dissertação (mestrado)
- ☐ Monografia (especialização)
- ☑ TCC (graduação)
- ☐ Artigo científico
- ☐ Capítulo de livro
- ☐ Livro
- ☐ Trabalho apresentado em evento
- ☐ Produto técnico e educacional - Tipo:

Nome completo do autor:
Heder Filho Silva Santos

Matrícula:
2022105231940018

Título do trabalho:
AVALIAÇÃO AUTOMÁTICA DE RESPOSTAS CURTAS EM PORTUGUÊS BRASILEIRO: UM ESTUDO SOBRE MODELOS DE LINGUAGEM, ENGENHARIA DE PROMPT E CARACTERÍSTICAS TEXTUAIS

RESTRIÇÕES DE ACESSO AO DOCUMENTO

Documento confidencial: ☑ Não ☐ Sim, justifique:

Informe a data que poderá ser disponibilizado no RIIF Goiano: 29 / 01 / 2026

O documento está sujeito a registro de patente? ☐ Sim ☑ Não

O documento pode vir a ser publicado como livro? ☐ Sim ☑ Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O(a) referido(a) autor(a) declara:

• Que o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;

• Que obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autoria, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;

• Que cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Iporá - GO
Local

29 / 01 / 2026
Data

gov.br Documento assinado digitalmente
HEDER FILHO SILVA SANTOS
Data: 29/01/2026 16:26:33-0300
Verifique em https://validar.iti.gov.br

Assinatura do autor e/ou detentor dos direitos autorais

Ciente e de acordo:

gov.br Documento assinado digitalmente
CLEON XAVIER PEREIRA JUNIOR
Data: 02/02/2026 15:53:48-0300
Verifique em https://validar.iti.gov.br

Assinatura do(a) orientador(a)

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema Integrado de Bibliotecas do IF Goiano - SIBi

Filho Silva Santos, Heder
S237 AVALIAÇÃO AUTOMÁTICA DE RESPOSTAS CURTAS EM PORTUGUÊS BRASILEIRO: UM ESTUDO SOBRE MODELOS DE LINGUAGEM, ENGENHARIA DE PROMPT E CARACTERÍSTICAS TEXTUAIS / Heder Filho Silva Santos. Iporá 2025.

36f. il.

Orientador: Prof. Dr. Cleon Xavier Pereira Junior.
Coorientador: Prof. Dr. Luiz Rodrigues.
Tcc (Bacharel) - Instituto Federal Goiano, curso de 0523194 - Bacharelado em Ciência da Computação - Iporá - 2020/1 (Campus Iporá).
I. Título.

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

Ata nº 80/2025 - GE-IP/CMPIPR/IFGOIANO

ATA DA SESSÃO DE JULGAMENTO DO TRABALHO DE CURSO
DE HEDER FILHO SILVA SANTOS

Aos três dias do mês de dezembro de dois mil e vinte e cinco, às quatorze horas e oito minutos, na sala de multimeios do bloco 2 do Instituto Federal Goiano – Campus Iporá, reuniu-se, em sessão pública, a banca examinadora designada na forma regimental pela Coordenação do Curso para julgar o trabalho de curso intitulado **"Avaliação automática de respostas curtas em português brasileiro: um estudo sobre modelos de linguagem, engenharia de prompt e características textuais"**, apresentado pelo acadêmico **Heder Filho Silva Santos** como parte dos requisitos necessários à obtenção do grau de Bacharel em Ciência da Computação. A banca examinadora foi presidida pelo orientador do trabalho de curso, Professor Doutor Cleon Xavier Pereira Júnior, tendo como membros a Professora Mestra Lais Candido Rodrigues da Silva Lopes e o professor doutor Marcos Alves Vieira. Aberta a sessão, o acadêmico expôs seu trabalho. Em seguida, foi arguido pelos membros da banca e:

(**X**) tendo demonstrado suficiência de conhecimento e capacidade de sistematização do tema de seu trabalho de curso, a banca conclui pela **aprovação** do acadêmico, sem restrições.

( ) tendo demonstrado suficiência de conhecimento e capacidade de sistematização do tema de seu trabalho de curso, a banca conclui pela **aprovação** do acadêmico, **condicionada a satisfazer as exigências** listadas na Folha de Modificação de Trabalho de Curso anexa à presente ata, no prazo máximo de 60 dias, a contar da presente data, ficando o professor orientador responsável por atestar o cumprimento dessas exigências.

( ) não tendo demonstrado suficiência de conhecimento e capacidade de sistematização do tema de seu trabalho de curso, a banca conclui pela **reprovação** do acadêmico.

Conforme avaliação individual de cada membro da banca, será atribuída a nota **9,5 (nove vírgula cinco)** para fins de registro em histórico acadêmico.

Os trabalhos foram encerrados às quinze horas e seis minutos. Nos termos do Regulamento do Trabalho de Curso do Bacharelado em Ciência da Computação do Instituto Federal Goiano – Campus Iporá, lavrou-se a presente ata que, lida e julgada conforme, segue assinada pelos membros da banca examinadora.

(Assinado Eletronicamente)

Prof. Dr. Cleon Xavier Pereira Júnior

(Assinado Eletronicamente)

Prof. Ma. Lais Cândido Rodrigues da Silva

(Assinado Eletronicamente)

Prof. Dr. Marcos Alves Vieira

Documento assinado eletronicamente por:

- **Cleon Xavier Pereira Junior**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 04/12/2025 07:19:13.
- **Lais Candido Rodrigues da Silva Lopes**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 04/12/2025 23:07:02.
- **Marcos Alves Vieira**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 08/12/2025 20:49:59.

Este documento foi emitido pelo SUAP em 03/12/2025. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 770236
Código de Autenticação: a7e810c72b

INSTITUTO FEDERAL GOIANO
Campus Iporá
Av. Oeste, Parque União, 350, Parque União, IPORA / GO, CEP 76.200-000
(64) 3674-0400

AGRADECIMENTOS

Registro minha profunda gratidão aos meus orientadores, Prof. Cleon Xavier Pereira Junior e Luiz Rodrigues, pela dedicação, apoio constante e pela orientação construída com paciência, confiança e rigor científica. Suas contribuições foram essenciais para que este trabalho se concretizasse.

Estendo meus agradecimentos a todos os professores que fizeram parte da minha formação. Cada aula, conselho e incentivo contribuíram para ampliar minha visão de mundo e fortalecer meu caminho na área da Computação.

Aos meus familiares, pelo incentivo constante, apoio incondicional e compreensão nos momentos de ausência dedicados aos estudos.

Agradeço também aos meus colegas, que compartilharam comigo desafios, aprendizados, conversas e risadas ao longo da graduação. A caminhada foi mais leve graças à companhia de vocês.

Meu carinho e agradecimento especial à minha namorada, pelo apoio, pela paciência e por acreditar no meu potencial em todos os momentos. Sua presença foi fundamental para eu chegar até aqui.

Por fim, agradeço a mim mesmo, pela persistência diante das dificuldades, pela dedicação que tornou este trabalho possível. Reconheço, com orgulho, o meu próprio esforço nesta trajetória.

RESUMO

SILVA SANTOS, HEDER FILHO. **AVALIAÇÃO AUTOMÁTICA DE RESPOSTAS CURTAS EM PORTUGUÊS BRASILEIRO: UM ESTUDO SOBRE MODELOS DE LINGUAGEM, ENGENHARIA DE PROMPT E CARACTERÍSTICAS TEXTUAIS.** dezembro, 2025. 25 f. Mografia – (Curso de Bacharel em Ciência da Computação), Instituto Federal Goiano - Campus iporá. Iporá, GO.

A Correção Automática de Respostas Curtas (em inglês, *Automatic Short Answer Grading* - ASAG) tem se destacado como alternativa promissora para reduzir o esforço humano em avaliações educacionais, embora ainda existam poucas investigações voltadas ao português brasileiro. Este estudo analisa o desempenho de três Modelos de Linguagem de Grande Escala (GPT-4o-mini, Sabiazinho-3 e Gemini 2.0-Flash) na tarefa de ASAG, avaliando todas as 128 combinações possíveis de sete componentes de engenharia de prompt e examinando como características textuais das respostas como número de palavras e riqueza lexical, influenciam o desempenho dos modelos. Os resultados indicam que a combinação de exemplos few-shot com rubrica explícita foi a mais eficaz, enquanto o raciocínio passo a passo beneficiou especialmente o GPT-4o-mini. O Sabiazinho-3 apresentou a maior concordância com avaliadores humanos, o Gemini 2.0-Flash obteve o menor erro médio absoluto, embora com alta taxa de alucinações, e o GPT-4o-mini produziu as saídas numéricas mais estáveis. Por fim, verificou-se que o perfil lexical das respostas impacta significativamente a qualidade da avaliação automática, sendo a faixa de riqueza lexical média a mais desafiadora para todos os modelos.

Palavras-chave: Correção Automática de Respostas Curtas; Modelos de Linguagem de Grande Escala; Engenharia de Prompt; Português Brasileiro.

ABSTRACT

SILVA SANTOS, HEDER FILHO. Automatic Short Answer Grading in Brazilian Portuguese: A Study on Language Models, Prompt Engineering, and Textual Characteristics. dezembro, 2025. 25 f. Trabalho de Conclusão de Curso – Bacharel em Ciência da Computação, Instituto Federal Goiano - Campus iporá. Iporá, GO, dezembro, 2025.

Automatic Short Answer Grading (ASAG) has emerged as a promising approach to reducing human effort in large-scale educational assessments, but studies focused on Brazilian Portuguese remain limited. This work evaluates the performance of three Large Language Models (GPT-4o-mini, Sabiazinho-3, and Gemini 2.0-Flash) in ASAG, testing all 128 possible combinations of seven prompt engineering components and examining how textual characteristics—such as word count and lexical richness—affect model accuracy. Results show that combining few-shot examples with explicit rubrics was the most effective strategy, while step-by-step reasoning particularly benefited GPT-4o-mini. Sabiazinho-3 achieved the highest agreement with human evaluators, Gemini 2.0-Flash obtained the lowest mean absolute error but exhibited a high hallucination rate, and GPT-4o-mini produced the cleanest and most consistent numeric outputs. Furthermore, the lexical profile of student responses significantly influenced model performance, with medium levels of lexical richness posing the greatest challenge across all models.

Keywords: Automatic Short Answer Grading; Large Language Models; Prompt Engineering; Brazilian Portuguese.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Avaliar respostas de estudantes é uma tarefa central em contextos educacionais, pois fornece um importante mecanismo de *feedback* que influencia diretamente o processo de aprendizagem (BURROWS; GUREVYCH; STEIN, 2015). No entanto, a correção manual de questões discursivas de resposta curta demanda tempo e esforço significativos por parte dos educadores (ELNAKA et al., 2021). Esse desafio se intensifica em ambientes com grande quantidade de participantes, como Cursos Online Abertos e Massivos (MOOCs), sistemas de ensino a distância e avaliações em larga escala, nos quais a agilidade e a consistência na correção são essenciais, mas difíceis de alcançar devido ao volume de respostas (PIECH et al., 2013).

A Correção Automática de Respostas Curtas (ASAG, do inglês *Automatic Short Answer Grading*) emerge como uma alternativa promissora para mitigar esses problemas. Essa abordagem utiliza métodos computacionais para analisar respostas textuais, estimar sua equivalência semântica em relação ao gabarito e atribuir notas automaticamente (SüZEN et al., 2020). Além de reduzir a carga de trabalho docente, ASAG pode contribuir para maior padronização, imparcialidade e rapidez no processo avaliativo (BURROWS; GUREVYCH; STEIN, 2015).

A correção manual, além de consumir tempo, está sujeita à subjetividade e à variabilidade entre avaliadores, o que pode comprometer a justiça e a consistência do processo avaliativo (BURROWS; GUREVYCH; STEIN, 2015). A ASAG não apenas promete escalabilidade para grandes volumes de dados, mas também oferece o potencial de um padrão de avaliação mais consistente, mitigando vieses humanos ao aplicar critérios pré-definidos de forma uniforme em todas as respostas (BONTHU; SREE; PRASAD, 2021). A adoção de ferramentas como a ASAG é, portanto, uma necessidade crescente para garantir a qualidade e a equidade em sistemas educacionais massivos.

Ao longo das últimas décadas, a evolução da Correção Automática de Respostas Curtas foi impulsionada sobretudo por modelos tradicionais de Processamento de Linguagem Natural (PLN), que dominaram a área antes do surgimento dos modelos de linguagem de grande escala (BURROWS; GUREVYCH; STEIN, 2015). Abordagens baseadas em extração de características como TF-IDF, *Bag-of-Words* e medidas de similaridade semântica aliadas a algoritmos supervisionados, como Máquinas de Vetores de Suporte (SVM), Regressão Linear e *Naive Bayes*, foram amplamente exploradas para estimar a correspondência entre a resposta do estudante e a resposta esperada (MELLO et al., 2025; LIU; KUSNER; BLUNSOM, 2020).

Esses métodos apresentaram avanços importantes ao oferecer avaliações consistentes em cenários controlados, especialmente quando combinados a técnicas de pré-processamento e engenharia manual de atributos. Contudo, apesar de sua relevância histórica e da robustez em tarefas específicas, tais modelos dependiam fortemente de representações superficiais do texto e frequentemente falhavam em capturar nuances semânticas essenciais.

Nesse cenário, os Modelos de Linguagem de Grande Escala (LLMs, do inglês *Large Language Models*) têm se destacado como ferramentas poderosas para tarefas de processamento de linguagem natural. Esses modelos, treinados em grandes corpora textuais, são capazes de capturar padrões semânticos e estruturais complexos (ZHUANG et al., 2023). Diferentemente de abordagens anteriores, baseadas em aprendizado por transferência ou engenharia manual de características, os LLMs frequentemente generalizam melhor entre

tarefas distintas e conseguem interpretar nuances linguísticas com maior profundidade (QIN et al., 2024). Estudos recentes sugerem, contudo, que seu desempenho pode ser ampliado quando associados a técnicas como ajuste fino e engenharia de *prompts*, que adaptam modelos pré-treinados para tarefas ou domínios específicos (WEI et al., 2021; CARPENTER et al., 2024).

Apesar dessas capacidades avançadas, o uso de LLMs em ASAG também introduz novos desafios que precisam ser cuidadosamente considerados. Embora esses modelos consigam interpretar nuances linguísticas de forma mais profunda que abordagens anteriores, eles podem apresentar variabilidade nas decisões, sensibilidade à formulação do *prompt* e tendência à geração de respostas alucinatórias fatores que podem comprometer a confiabilidade necessária em contextos educacionais (YAN et al., 2024; GRéVISSE, 2024). Além disso, aspectos como custo computacional, necessidade de controle sobre o formato da saída e diferenças entre idiomas frequentemente influenciam seu desempenho (BANG et al., 2023). Assim, a adoção de LLMs para fins avaliativos requer não apenas avanços metodológicos, mas também uma compreensão sistemática de componentes de *prompt* e características textuais.

Pesquisas em língua inglesa têm explorado o potencial e os desafios dos LLMs no contexto de ASAG. Por exemplo, Chamieh, Zesch e Giebermann (2024) compararam modelos da família GPT e LLaMA a abordagens supervisionadas tradicionais em cenários *zero-shot* e *few-shot*, observando desempenho limitado em perguntas que exigiam raciocínio complexo ou conhecimento especializado. Por outro lado, Grévisse (2024) avaliaram o GPT-4 e o Gemini 1.0 em um contexto educacional real, corrigindo 2.288 respostas escritas em três idiomas diferentes. Apesar desses avanços, persistem lacunas importantes, como a predominância de estudos em língua inglesa, a ausência de investigações sistemáticas sobre o impacto de diferentes componentes de *prompt* e a escassez de análises qualitativas de erros.

No contexto do português brasileiro, as contribuições ainda são incipientes. Mello et al. (2025) apresentam uma análise comparativa entre modelos tradicionais e o GPT-4, destacando o papel central da engenharia de *prompts* na melhoria do desempenho do modelo. De maneira complementar, Mello et al. (2024) avaliaram sistematicamente 128 combinações de *prompt*, evidenciando a relevância de elementos como estratégias *few-shot* e justificativas no aumento da eficácia do ASAG em português.

Diante desse cenário, este trabalho investiga o uso de LLMs para ASAG especificamente no português brasileiro. São avaliados os modelos GPT4o-mini, Sabiazinho-3 e Gemini 2.0-Flash, analisando sua capacidade de interpretar e classificar respostas textuais curtas. Conduzimos experimentos envolvendo diferentes técnicas de engenharia de *prompt*, além de investigar o efeito de componentes específicos dos *prompts* sobre a acurácia dos modelos. O objetivo é comparar o desempenho entre arquiteturas diversas e identificar quais combinações de modelos e estratégias de *prompt* produzem resultados mais robustos e precisos para o contexto educacional brasileiro.

O restante deste trabalho está estruturado da seguinte forma: o Capítulo 2 apresenta os conceitos fundamentais sobre Correção Automática de Respostas Curtas e a evolução dos Modelos de Linguagem de Grande Escala. O Capítulo 3 detalha a metodologia de pesquisa, incluindo a descrição do *dataset* em Português Brasileiro, os modelos de LLMs selecionados e as estratégias de engenharia de *prompt* aplicadas. Em seguida, os Capítulos 4 e 5 apresentam os resultados quantitativos da avaliação dos modelos e a análise qualitativa dos erros. Finalmente, o Capítulo 6 resume as principais conclusões deste estudo, suas implicações e sugestões para trabalhos futuros na área.

2 FUNDAMENTAÇÃO TEÓRICA

Este estudo visa 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, este capítulo apresenta uma contextualização geral do ASAG e da engenharia de *prompts*. Em seguida, o capítulo apresenta a fundamentação teórica da área de classificação automática de respostas curtas. Finalmente, apresentamos as questões de pesquisa que norteiam este estudo.

2.1 APRENDIZADO DE MÁQUINA

O aprendizado de máquina (AM) é uma da área da inteligência Artificial(IA) importante para sistemas de predição de padrões e tomada de decisões com base em dados estabelecidos (MITCHELL, 1997). Assim, seu impacto é vasto e tem sido utilizado em diversas áreas, como saúde, finanças e educação.

2.1.1 TIPOS DE APRENDIZADO DE MÁQUINA

Aprendizado supervisionado: Os algoritmos aprendem a partir de um conjunto de treinamento. Este conjunto possui entradas e saídas corretas. Assim, os algoritmos aprendem ao repassar estes dados, onde medem seus erros com uma função de perda, se ajustando de modo a otimizar sua resposta, minimizando a margem de erro. Exemplos comuns de utilização são em modelos de classificação e regressão (BISHOP, 2006).

Aprendizado não supervisionado: Os algoritmos recebem um conjunto de treinamento não rotulado, onde buscam por padrões ocultos e agrupamento de dados. Tem um caráter exploratório, se tornando ideal para exploração de dados. Exemplos de seu uso são as regras de associação e a clusterização (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Aprendizado auto-supervisionado: Este tipo de aprendizado utiliza uma abordagem, onde os dados de treinamento são rotulados a partir das entradas, criando rótulos implícitos. Ela pode ser aplicada quando é necessário usar um aprendizado não supervisionado, porém, é preciso usar um aprendizado supervisionado. É uma técnica emergente, que vem sendo utilizada em visão computacional (DEVLIN et al., 2019).

Aprendizado por reforço No aprendizado por reforço, um agente interage com um ambiente dinâmico, tomando ações e recebendo recompensas ou penalidades como *feedback*. O objetivo do agente é aprender uma política de ações que maximize sua recompensa acumulada ao longo do tempo. Esse tipo de aprendizado é amplamente utilizado em robótica, jogos e otimização de processos, sendo aplicado, por exemplo, em algoritmos como *Q-Learning* (SUTTON; BARTO, 2018).

Um fator de importância para se considerar, ao se trabalhar com o aprendizado de máquina é se ter um conjunto de dados robusto, porém não é sempre que se tem um conjunto de dados estáticos de amostras rotuladas. Assim, a fim de minimizar a necessidade de um grande e robusto conjunto de dados, o aprendizado de máquina ativa surge como uma grande alternativa para o treinamento de modelos (OLSSON, 2009).

O Aprendizado de Máquina ativo é uma abordagem na qual o modelo identifica quais instâncias não rotuladas são mais relevantes para receber anotações de especialistas,

reduzindo o esforço de rotulagem. Esse processo visa aumentar a qualidade dos resultados (SETTLES, 2009). No contexto de classificação textual, é uma técnica importante, pois, diante de grandes volumes de texto, pode ser inviável rotular manualmente todos os documentos (NOVAK; MLADENIČ; GROBELNIK, 2006).

Desta maneira, cada abordagem de aprendizado de máquina desempenha um papel crucial na classificação textual, variando conforme a disponibilidade de dados rotulados, a necessidade de exploração e a adaptação dinâmica do modelo às mudanças no ambiente (BONTHU; SRIPADA; PRASAD, 2021). Isso ocorre porque cada técnica, seja supervisionada, não supervisionada ou de reforço, oferece diferentes vantagens em termos de descoberta de padrões, robustez à falta de rótulos e aprendizagem contínua. Dessa forma, é possível alcançar maior eficiência na categorização de textos sob cenários diversos.

2.2 CLASSIFICAÇÃO TEXTUAL

A classificação textual configura-se como uma das aplicações mais relevantes do processamento de linguagem natural (PLN), tendo por objetivo organizar e categorizar documentos textuais em classes predefinidas com base em suas características semânticas e sintáticas. Essa técnica, que se apoia em modelos estatísticos e algoritmos de aprendizado de máquina, permite a automatização da análise de grandes volumes de dados, facilitando o gerenciamento da informação e a extração de conhecimento a partir de textos. O desenvolvimento de abordagens robustas para a classificação textual é crucial para a implementação de sistemas de recomendação, filtragem de *spam*, análise de sentimentos, entre outras aplicações, contribuindo de maneira significativa para a eficiência e eficácia dos processos de tomada de decisão em ambientes complexos.

2.2.1 MÉTODOS DE CLASSIFICAÇÃO TEXTUAL

A classificação textual é uma tarefa central no PLN e envolve a atribuição de categorias ou rótulos a textos com base em seu conteúdo. Para realizar essa tarefa, diversas abordagens e técnicas têm sido desenvolvidas ao longo dos anos, tanto no campo dos métodos estatísticos clássicos quanto nos métodos baseados em modelos de linguagem de Grande escala (LLMs).

2.2.1.1 PRÉ-PROCESSAMENTO DOS DADOS

Antes da aplicação de qualquer modelo de classificação, o texto precisa ser convertido em uma forma que possa ser processada computacionalmente (MANNING; SCHüTZE, 1999). As etapas de pré-processamento incluem:

1. Tokenização: Dividir o texto em unidades menores (*tokens*), que podem ser palavras, frases ou caracteres.
2. Normalização: Inclui a conversão para letras minúsculas, remoção de pontuação, *stopwords* (palavras de baixa relevância) e, em alguns casos, a aplicação de técnicas de lematização.

A fim de obter uma representação vetorial, pode-se citar técnicas como o *Bag-of-Words* (BoW), que representa textos como um vetor de frequências de palavras, ignorando a ordem. Também é possivel utilizar o *Term Frequency-Inverse Document Frequency*(TF-IDF), que pondera a frequência das palavras considerando a importância relativa delas em um corpus

Superando outros métodos de representação vetorial, é notório destacar os *word embeddings*. Eles representam uma técnica fundamental no processamento de linguagem natural, onde palavras são convertidas em vetores de números reais em um espaço contínuo. Essa abordagem, exemplificada por modelos como Word2Vec, GloVe e FastText, permite que termos semanticamente próximos sejam representados por vetores similares, facilitando a captura de relações contextuais e semânticas entre palavras.

2.2.1.2 MODELOS TRADICIONAIS

Os modelos tradicionais de classificação textual se caracterizam pela sua abordagem robusta. Esses modelos adotam estratégias que facilitam a separação dos dados em categorias distintas, ainda que existam desafios impostos pela alta dimensionalidade dos textos. Ao explorar técnicas baseadas em princípios estatísticos e algoritmos de aprendizado, os métodos demonstram versatilidade e adaptabilidade (SEBASTIANI, 2002; MELLO et al., 2025).

Naive Bayes: Baseado no teorema de Bayes, assume a independência entre as características. É rápido e eficaz para muitos problemas de classificação de texto, apesar da simplificação assumida.

Máquinas de Vetores de Suporte (SVM): Encontram um hiperplano que separa as classes de maneira ótima. SVMs são especialmente úteis quando os dados são de alta dimensionalidade, como em representações BoW ou TF-IDF.

Árvores de Decisão e *Random Forests*: Utilizam critérios de divisão para particionar os dados em grupos homogêneos. Podem ser aplicadas à classificação textual, embora muitas vezes necessitem de um bom pré-processamento para lidar com a alta dimensionalidade.

2.2.1.3 MODELOS DE *DEEP LEARNING*

O surgimento e o avanço das técnicas de *deep learning* transformaram radicalmente a abordagem para a classificação textual, elevando a capacidade de extrair representações profundas dos dados (BONTHU; SRIPADA; PRASAD, 2021). Ao invés de depender exclusivamente de características extraídas manualmente ou de representações simplificadas, as arquiteturas baseadas em redes neurais permitem o aprendizado de padrões complexos e relações semânticas intrincadas a partir dos próprios dados, enriquecendo a compreensão dos contextos linguístico (MALEKZADEH et al., 2021).

Ao empregar estruturas de rede sofisticadas e técnicas avançadas de treinamento, os modelos de *deep learning* mostram-se particularmente eficazes em lidar com as sutilezas da linguagem natural, capturando nuances contextuais e variações fundamentais para uma categorização mais precisa e robusta dos textos.

Redes Neurais Embora originalmente desenvolvidas para dados visuais, elas se aplicam com sucesso à classificação de texto ao identificar padrões locais nas sequências textuais (MOTA et al., 2020).

Redes Neurais Profundas Modelos como LSTM (Long Short-Term Memory) e GRU (Gated Recurrent Units) conseguem capturar dependências sequenciais, importantes para entender a estrutura e o contexto do texto (PERUMAL et al., 2024).

2.2.1.4 MODELOS DE LINGUAGEM DE GRANDE ESCALA

Os LLMs representam um avanço significativo no campo do 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; 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; JAIN; KANKANHALLI, 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).Tais limitações reforçam a necessidade de uma abordagem sistemática de engenharia de *prompt* e validação das saídas, visando mitigar erros e viabilizar o uso seguro e eficaz dos LLMs em contextos avaliativos.

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 (SANTU; 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 (SANTU; 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). Juntas, essas abordagens elevam a precisão e a qualidade das respostas em cenários variados, de questões aritméticas a avaliações educacionais.

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 e 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; 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 CORREÇÃO AUTOMÁTICA DE RESPOSTAS CURTAS

A Correção Automática de Respostas Curtas (ASAG, do inglês *Automatic Short Answer Grading*) consiste em utilizar métodos de PLN e inteligência artificial para avaliar respostas discursivas breves de estudantes de forma automática (BURROWS; GUREVYCH; STEIN, 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; PURNAMASARI; RATNA, 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 e 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; JOORABCHI; HAYES, 2022). Diversos estudos exploraram arquiteturas de redes neurais para ASAG. Camus e Filighera (2020) investigaram o uso de transformadores utilizando o modelo BERT para pontuação automática, enquanto Sung, Dhamecha e Mukhi (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 brasileiro (GALHARDI; SOUZA; BRANCHER, 2020). A grande parte dos estudos e do desenvolvimento na área de ASAG concentra-se no idioma inglês (BURROWS; GUREVYCH; STEIN, 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).

2.5 QUESTÕES DE PESQUISA

A investigação proposta articula um conjunto de questões de pesquisa que visam compreender, de maneira sistemática, como diferentes fatores — arquiteturas de modelos, engenharia de *prompt* e características textuais das respostas — influenciam o desempenho de LLMs na tarefa de avaliação automática de respostas curtas (ASAG) em português brasileiro. Embora trabalhos recentes comparem modelos como GPT-3.5, GPT-4 e versões open-source em múltiplos idiomas (CHANG; GINTER, 2024), ainda são escassas as análises centradas especificamente no português brasileiro, e as existentes permanecem limitadas principalmente aos modelos da família GPT (MELLO et al., 2024). Essa lacuna indica a necessidade de ampliar a investigação para modelos com arquiteturas distintas, incluindo aqueles treinados nativamente em português, bem como de compreender como aspectos textuais das respostas e componentes do *prompt* influenciam o desempenho dos sistemas.

Nesse contexto, foram formuladas as seguintes Questões de Pesquisa (QPs):

QP1: *Qual é o desempenho de diferentes modelos de LLMs no contexto de ASAG para o português brasileiro?*

QP2: *Quão suscetíveis esses modelos são à geração de alucinações durante a tarefa de ASAG?*

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

QP4: *Como características textuais influenciam o desempenho de LLMs na ASAG?*

No presente estudo, o Experimento 1 aborda diretamente as QPs 1, 2 e 3, ao comparar modelos com diferentes tamanhos e arquiteturas em um cenário controlado de atribuição de notas a respostas dissertativas curtas em português brasileiro. A partir desses resultados, buscou-se estabelecer evidências empíricas que fundamentem recomendações tanto de modelos quanto de estratégias de *prompting* adequadas para aplicações educacionais de ASAG.

Enquanto o Experimento 2 aborda a QP4, concentrando-se na relação entre as propriedades textuais das respostas estudantis e a precisão das avaliações automáticas realizadas pelos modelos. Nesta etapa, o foco desloca-se da comparação entre arquiteturas e estratégias de *prompt* para a análise da própria resposta.

3 METODOLOGIA

Este trabalho segue a metodologia apresentada por Mello et al. (2024) para examinar, de forma sistemática, o impacto de componentes de engenharia de *prompt* na tarefa de ASAG em português brasileiro. Porém, esta pesquisa se difere ao analisar não somente os elementos de engenharia de *prompt* em modelos GPT, como também comparar o desempenho de diferentes modelos de LLM, incluindo modelo treinado para o português.

Com base nisso, este capítulo apresenta o conjunto de dados utilizado para responder às questões de pesquisa, aplicando o contexto da língua portuguesa. Em seguida, com base na revisão da literatura, foram levantados os componentes relevantes na construção de *prompts* que serão analisados nesta pesquisa, seguidos pelos modelos escolhidos para a aplicação do experimento. Por fim, serão apresentados como os resultados devem ser avaliados.

3.1 DATASET

O conjunto de dados utilizado neste estudo, denominado PT_ASAG, foi proposto por Galhardi, Souza e Brancher (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 (MELLO et al., 2025).

Assim como em Mello et al. (2024), foram aplicados aproximadamente 30% do volume de dados original, correspondendo a cerca de 2.641 respostas (Tabela 1). Esse subconjunto foi selecionado de forma aleatória, garantindo que todas as questões permanecessem representadas e reduzindo potenciais vieses de seleção. Tal 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, foram selecionados 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

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

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).

- **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; 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. A temperatura é um parâmetro que controla o nível de aleatoriedade na geração das respostas: valores próximos de 0 tornam o modelo mais determinístico e previsível, enquanto valores mais altos favorecem diversidade, criatividade e variação nas saídas. Como o objetivo neste estudo era obter apenas a nota atribuída de forma consistente, optou-se por manter *temperature* = 0,0 nos modelos que permitem ajustes diretos, reduzindo o risco de variação indesejada entre execuções.

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). Foram aderidos também as medidas e práticas recomendadas na literatura (WHITE et al., 2023). Primeiramente, foram definidos 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.

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 por meio de três métricas complementares: o Coeficiente de Concordância de Cohen (κ), o Erro Médio Absoluto (MAE) e o Erro Quadrático Médio (RMSE). Essa combinação permite uma análise balanceada, contemplando tanto a consistência qualitativa (acordo) quanto a precisão quantitativa (erro).

O κ quantifica o grau de concordância entre as notas geradas pelos modelos e aquelas atribuídas por avaliadores humanos, ajustando o acordo pela chance aleatória e oferecendo uma medida robusta de confiabilidade (FLEISS; COHEN; EVERITT, 1969). O MAE expressa a discrepância média absoluta entre a nota prevista e a nota de referência, fornecendo uma interpretação direta do erro médio (ZHAO et al., 2017). 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 (ZHAO et al., 2017).

Além dessas métricas principais, as análises do Experimento 2 utilizaram o erro médio simples como indicador complementar para investigar a relação entre o desempenho dos LLMs e características textuais das respostas, como extensão, riqueza lexical e presença de palavras-chave.

4 EXPERIMENTO 1: APLICAÇÃO DOS MODELOS

Esta seção pretende analisar de forma sistemática o desempenho de diferentes Modelos de Linguagem de Grande Escala na tarefa de Correção Automática de Respostas Curtas em português brasileiro. Para isso, foram aplicadas técnicas de engenharia de *prompt* que permitiram não apenas avaliar a eficácia dos modelos em reproduzir julgamentos humanos, mas também identificar os componentes de *prompt* que mais impactam a qualidade das avaliações. Além da comparação entre distintas arquiteturas de LLMs, buscou-se investigar sua suscetibilidade à geração de alucinações e sua capacidade de oferecer resultados consistentes em métricas qualitativas e quantitativas. Dessa forma, este experimento estabelece a base para compreender como diferentes estratégias de *prompting* podem otimizar o uso de LLMs no contexto educacional brasileiro.

4.1 MÉTRICAS DO EXPERIMENTO

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 somente 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 QP2, 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; SHETH; DAS, 2023; XU; JAIN; KANKANHALLI, 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*.

4.2 RESULTADOS

4.2.1 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 (4). 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.

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

Por sua vez, o Gemini 2.0-Flash aproxima-se do desempenho do sabiazinho-3, alcançando $\kappa = 0{,}49$ 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.

4.2.2 QP2: QUÃO SUSCETÍVEIS ESSES MESMOS MODELOS SÃO À GERAÇÃO DE ALUCINAÇÕES DURANTE A TAREFA DE ASAG?

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 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

4.2.3 QP3: QUAIS COMPONENTES ESPECÍFICOS DO DESIGN DE *PROMPT* PODEMD 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

4.3 DISCUSSÃO

Os resultados obtidos neste experimento 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 experimento 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 experimento 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; SOUZA; BRANCHER, 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).

A taxa de alucinação emerge, portanto, como indicador estruturante de confiabilidade nos fluxos de correção automática em larga escala. O GPT-4o-mini apresenta média de 0,76% de linhas alucinatórias por *prompt*, sem cortes, evidenciando que modelos de maior porte, embora mais onerosos, oferecem uma base textual segura e minimizam o pós-processamento. O Sabiázinho exibe média de 8,30% (17 combinações descartadas), permanecendo abaixo do limiar de 25% definido neste estudo; isso indica um bom equilíbrio entre custo e precisão, desde que mecanismos simples de validação bloqueiem casos pontuais. Em contraste, o Gemini 2.0-Flash alcança 48,50% de alucinações (76 linhas removidas), inviabilizando uma automação integral, pois quase metade das saídas exigiria filtragem intensa antes da análise estatística.

5 EXPERIMENTO 2: ANALISE DESCRITIVA

Nesta seção, são apresentados os resultados obtidos a partir da análise do desempenho dos modelos de linguagem em função das características das respostas dos estudantes. O objetivo foi identificar como atributos específicos do perfil das respostas, como extensão textual, riqueza lexical, presença de palavras-chave e similaridade em relação ao gabarito, influenciam a qualidade das notas atribuídas pelos modelos.

5.1 MÉTRICAS DO EXPERIMENTO

O Experimento 2 teve como objetivo investigar como diferentes características textuais das respostas — tais como riqueza lexical (TTR), extensão (número de palavras), presença de palavras-chave e similaridade com o gabarito — influenciam o desempenho dos modelos de linguagem. Diferentemente das métricas globais do estudo (Cohen's κ, MAE e RMSE), empregadas para avaliar a concordância estrutural entre modelos e humanos, aqui adotamos métricas descritivas locais, adequadas para analisar o comportamento dos modelos em subgrupos específicos do conjunto de respostas.

A métrica central utilizada foi o erro médio simples, calculado como a diferença absoluta entre a nota atribuída pelo modelo e a nota de referência para cada resposta. Esse indicador permite identificar tendências de desempenho associadas a diferentes perfis textuais, funcionando como uma medida sensível a variações dentro de cada faixa analisada. Sua simplicidade facilita comparações entre modelos e entre categorias de respostas, tornando-o apropriado para análises exploratórias.

Além do erro médio, foram utilizadas também distribuições de erro em forma de boxplots, permitindo observar variabilidade, dispersão e presença de valores atípicos dentro de cada faixa de característica textual. Essa abordagem complementa o erro médio ao revelar não somente o valor típico do erro, mas também sua estabilidade, aspecto fundamental para compreender a robustez dos modelos frente a diferentes tipos de produção escrita.

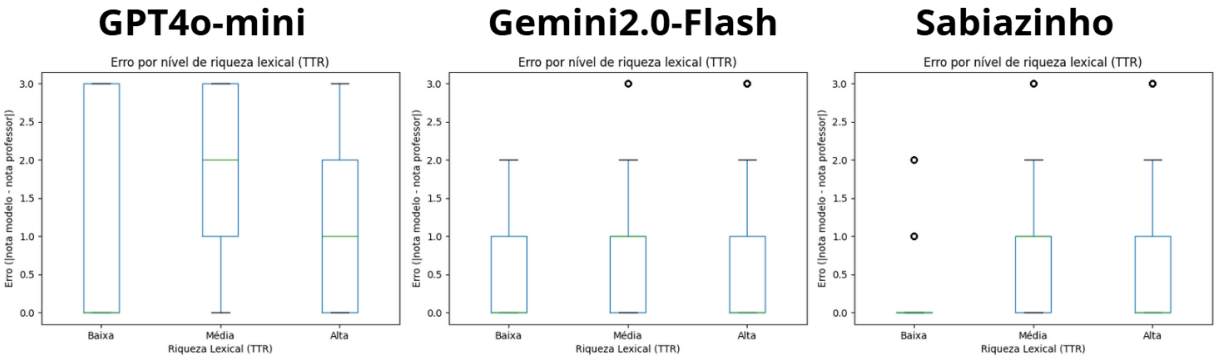

Figura 1 – Distribuição dos erros conforme a faixa de TTR, para cada modelo

Tabela 7 – Erro médio por faixa de TTR para cada modelo

Modelo	Baixo	Médio	Alta
GPT4o-mini	1.17	2.05	1.17
sabiazinho-3	0.17	0.78	0.48
Gemini 2.0-Flash	0.44	0.68	0.52

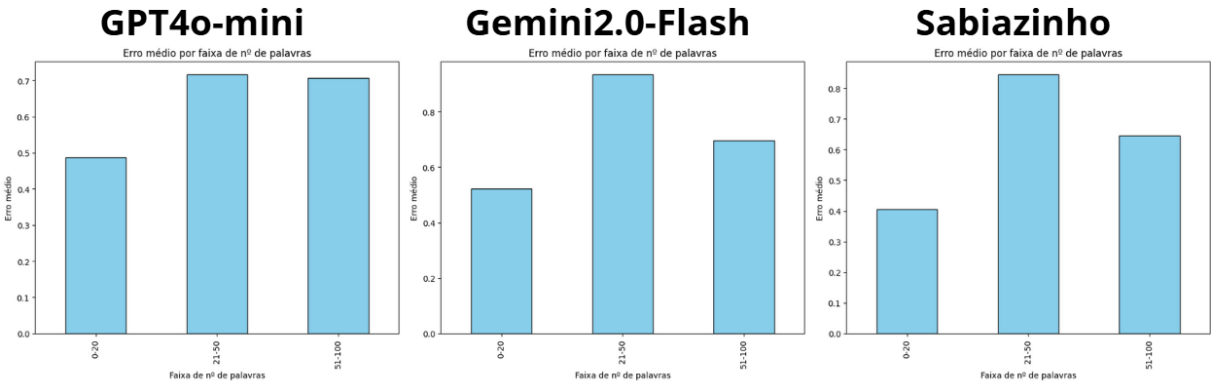

Figura 2 – Erro médio por faixa de número de palavras

5.2 QP4: COMO CARACTERÍSTICAS TEXTUAIS INFLUENCIAM O DESEMPENHO DE LLMS PARA A ASAG

A Figura 1 e a Tabela 7 evidenciam diferenças importantes no impacto da riqueza lexical sobre o desempenho dos modelos. O GPT-4o-mini apresentou o maior erro médio na faixa de TTR médio (2,05), acompanhado de grande dispersão, indicando sensibilidade a respostas com diversidade vocabular intermediária. O Sabiazinho-3, embora tenha alcançado o melhor desempenho absoluto em TTR baixo (0,17), também apresentou aumento expressivo do erro em TTR médio (0,78), sugerindo que respostas mais elaboradas, mas sem uso consistente de palavras-chave, representam um desafio. Já o Gemini 2.0-Flash demonstrou maior estabilidade entre as faixas, com erros próximos em todos os níveis (0,44, 0,68 e 0,52), confirmando sua robustez diante de diferentes perfis lexicais.

Comparativamente, é notório que os modelos apresentaram maior dificuldade na faixa de TTR médio, confirmando que respostas com diversidade lexical moderada tendem a ser mais ambíguas para a avaliação. O GPT-4o-mini destacou-se negativamente pela instabilidade e altos erros nessa categoria, enquanto o Gemini 2.0-Flash demonstrou maior robustez, mantendo erros relativamente estáveis entre as faixas. Já o Sabiazinho-3 apresentou os menores erros absolutos, sobretudo em respostas de TTR baixo, confirmando sua adequação ao português brasileiro. Esses resultados indicam que o perfil lexical da resposta é um fator determinante no desempenho dos modelos, e que arquiteturas distintas apresentam sensibilidades diferentes a essa característica.

A Figura 2 apresenta gráficos do erro médio por faixa de número de palavras para cada um dos modelos analisados. Nessa figura é evidente que os modelos apresentaram maior discrepância em respostas intermediárias (21–50 palavras) sugerindo que, nesse intervalo, é importante investigar de maneira mais detida a necessidade, ou não, de ajustes nos modelos, no sentido de mitigar tal situação. Em contrapartida, respostas curtas (até 20 palavras) foram mais bem avaliadas, especialmente pelo Sabiazinho-3, que obteve o menor erro médio nessa faixa. Já em respostas longas, observou-se redução do erro em relação ao grupo intermediário, indicando que a presença de mais detalhes textuais auxilia

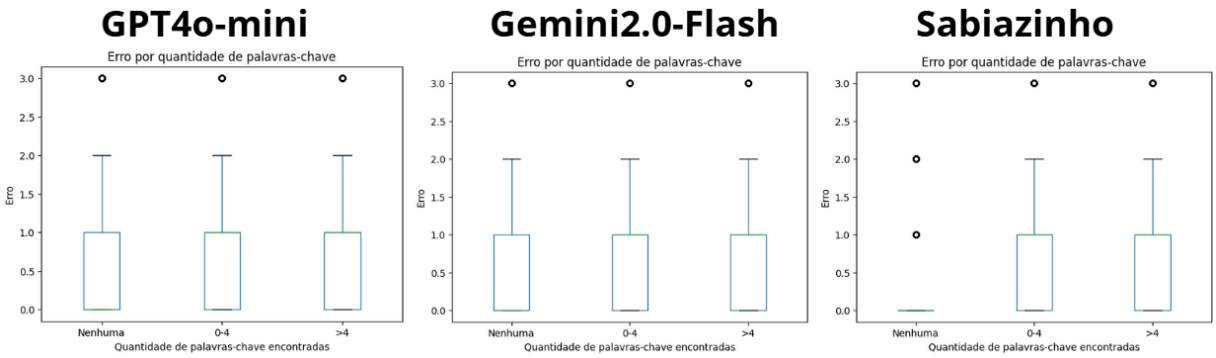

Figura 3 – Distribuição do erro médio por faixa de palavra-chave

os modelos na aproximação da nota atribuída por avaliadores humanos.

A análise da influência da presença de palavras-chave (3) revelou que a quantidade desses termos não impactou de maneira significativa o erro médio do GPT-4o-mini e do Gemini 2.0-Flash, cujas distribuições permaneceram estáveis em todas as faixas avaliadas. Por outro lado, o Sabiazinho-3 apresentou comportamento distinto: em respostas sem ocorrência de palavras-chave, obteve mediana de erro próxima a zero, mas com maior dispersão, enquanto nas demais categorias o erro médio se elevou consistentemente. Esses resultados sugerem que, enquanto modelos globais tendem a apoiar-se em aspectos semânticos mais amplos, o modelo nativo em português demonstra maior sensibilidade à presença ou ausência de termos-chave, o que pode tanto favorecer a precisão em respostas objetivas quanto gerar instabilidade em produções mais complexas.

5.3 DISCUSSÃO

Os resultados do Experimento 2 evidenciam que características textuais das respostas influenciam de maneira significativa o desempenho dos modelos na tarefa de ASAG. A riqueza lexical (TTR) revelou-se especialmente determinante: todos os modelos apresentaram maior dificuldade na faixa intermediária, indicando que respostas com diversidade lexical moderada tendem a ser mais ambíguas e menos alinhadas ao gabarito. O GPT-4o-mini foi o mais sensível a essa faixa, enquanto o Gemini 2.0-Flash mostrou maior estabilidade. O Sabiazinho-3 teve excelente desempenho em TTR baixo, coerente com seu ajuste ao português brasileiro, mas apresentou aumento de erro conforme a complexidade lexical cresceu.

A análise do número de palavras reforça esse padrão. Respostas curtas foram avaliadas com maior precisão, enquanto respostas intermediárias (21–50 palavras) geraram maior erro médio, sugerindo um "ponto crítico" onde a resposta inclui informações irrelevantes, mas não detalha suficientemente a ideia central. Nas respostas longas, o erro diminuiu, indicando que maior densidade semântica ajuda os modelos a identificar elementos relevantes. Já a presença de palavras-chave afetou os modelos desigualmente: GPT-4o-mini e Gemini 2.0-Flash mantiveram estabilidade, enquanto o Sabiazinho-3 mostrou maior variabilidade em respostas mais complexas.

Integradamente, os achados indicam que o perfil textual da resposta é um fator estrutural para o desempenho dos LLMs. Embora todos reproduzam tendências gerais de avaliação, suas limitações reforçam a necessidade de estratégias que reduzam a influência de variações estilísticas como ajustes de *prompt* ou validação semântica para garantir avaliações mais consistentes em contextos pedagógicos reais.

6 CONSIDERAÇÕES FINAIS

6.1 LIMITAÇÕES

Embora os resultados deste estudo ofereçam evidências promissoras sobre o uso de LLMs para ASAG em português brasileiro, algumas limitações precisam ser consideradas para uma interpretação adequada dos achados. A primeira delas diz respeito ao uso de aproximadamente 30% do conjunto original de dados, uma amostra selecionada de forma aleatória para viabilizar os experimentos em termos de tempo e custo computacional. Apesar de esse subconjunto superar em escala diversos estudos da literatura, ele permanece restrito ao domínio de Biologia do ensino médio, limitando a generalização dos resultados para outras áreas do conhecimento, faixas etárias ou estilos de escrita.

Outra limitação importante envolve a incidência de alucinações nos modelos. Em certas combinações de *prompts*, observamos taxas superiores a 25%, um fenômeno abundantemente relatado na literatura como uma limitação estrutural dos LLMs (XU; JAIN; KANKANHALLI, 2025). Embora o critério de corte adotado tenha mitigado a influência desses casos nas métricas globais, a presença de alucinações coloca em evidência a fragilidade dos modelos em cenários que exigem precisão e consistência. Tal comportamento pode comprometer a confiabilidade de sistemas automatizados de avaliação, especialmente quando usados sem supervisão humana.

Além disso, o presente estudo utilizou *prompts* essencialmente estáticos: cada LLM recebe a pergunta e devolve uma nota sem qualquer iteração adicional. Esse formato, embora adequado para fins de *benchmarking*, não captura a natureza dinâmica da prática docente, em que há espaço para argumentação, explicação e revisão. Em ambientes educacionais reais, estudantes tendem a justificar suas respostas ou reformulá-las após *feedback*, e essas interações podem influenciar diretamente a avaliação. A ausência dessa dimensão limita o realismo dos experimentos.

Por fim, a análise não contemplou diferenças internas entre perguntas, como níveis de dificuldade, granularidade conceitual ou ambiguidades linguísticas. Embora o *dataset* inclua múltiplos itens avaliativos, o estudo tratou todas as questões como equivalentes, o que pode ocultar variações importantes no desempenho dos LLMs conforme o tipo de habilidade cognitiva exigida.

6.2 TRABALHOS FUTUROS

Diante das limitações observadas, diversas oportunidades de investigação se apresentam para ampliar a compreensão sobre o uso de LLMs em ASAG no contexto do português brasileiro. Um primeiro caminho consiste em expandir o escopo dos dados utilizados, incorporando disciplinas adicionais, como Língua Portuguesa, Matemática e Ciências Humanas, bem como diferentes níveis de ensino. Essa ampliação permitiria testar a robustez dos modelos frente a variadas formas de expressão escrita e diferentes graus de complexidade conceitual.

Outra direção promissora envolve o aprofundamento do estudo das alucinações. Pesquisas futuras podem explorar a eficácia de estratégias como ajustes finos de parâmetros de inferência, *prompt engineering* mais sofisticado, filtros automáticos de validação semântica e abordagens híbridas que combinem modelos generativos com heurísticas linguísticas ou modelos discriminativos. Tais soluções podem reduzir significativamente a incidência de

respostas incorretas ou desconexas, aumentando a confiabilidade do sistema para aplicações educacionais.

Também se mostra relevante investigar cenários avaliativos mais interativos. Modelos *multi-turn*, em que o LLM revisa sua nota após explicações ou contra-argumentos do aluno, aproximam-se mais de práticas reais e podem oferecer avaliações mais transparentes e formativas. Esses experimentos abririam espaço para estudar como LLMs lidam com justificativas, raciocínio metacognitivo e instruções de refinamento, fatores essenciais em contextos pedagógicos.

Por fim, há potencial para comparar a engenharia de *prompt* com abordagens supervisionadas ou semissupervisionadas, como *fine-tuning* em corpora específicos de avaliação escolar ou *pipelines* de Recuperação Aumentada por Geração (RAG). Esses métodos podem melhorar a consistência das notas atribuídas, reduzir alucinações e adaptar o comportamento dos modelos às particularidades da escrita estudantil brasileira. Explorar essas alternativas permitirá delinear caminhos mais sólidos para sistemas de ASAG que sejam confiáveis, transparentes e adequados ao uso pedagógico.

6.3 CONCLUSÃO

Este trabalho investigou de forma sistemática o desempenho de Modelos de Linguagem de Grande Escala na tarefa de Correção Automática de Respostas Curtas em português brasileiro, com foco na influência da engenharia de *prompt*, da arquitetura dos modelos e das características textuais das respostas. Os resultados obtidos permitem responder de maneira consistente às quatro Questões de Pesquisa propostas, além de oferecer contribuições relevantes para o avanço da área no contexto nacional.

Primeiramente, verificou-se que o desempenho dos modelos varia de acordo com suas características arquiteturais e com a adaptação linguística a dados em português brasileiro. Entre os modelos avaliados, o sabiazinho-3, treinado majoritariamente em português, destacou-se com os melhores valores de κ (até 0,50) e bons índices de erro (MAE médio = 0,38). Esses resultados reforçam a vantagem de modelos nativamente adaptados ao idioma para tarefas educacionais que exigem análise semântica específica do português. O Gemini 2.0-Flash apresentou desempenho competitivo em termos de erro médio, porém com queda de consistência categórica. Já o GPT-4o-mini, embora não tenha liderado as metricas, demonstrou uma característica particularmente valiosa: foi o modelo mais robusto contra alucinações, com apenas 0,76% de ocorrências — fator essencial para automações em escala.

No tocante à engenharia de *prompt*, observou-se que a combinação de *few-shot* e rubrica constitui o núcleo das estratégias mais eficazes. Esses dois componentes estiveram presentes na maioria das melhores configurações entre todos os modelos, confirmando achados prévios da literatura e demonstrando que fornecer exemplos e critérios explícitos direciona o raciocínio do modelo e reduz interferências subjetivas(CARPENTER et al., 2024; MELLO et al., 2025).

Por fim, o Experimento 2 evidenciou que características textuais da resposta do estudante influenciam diretamente o desempenho dos modelos, especialmente no que diz respeito à riqueza lexical, extensão textual e presença de palavras-chave. Em geral, o sabiazinho-3 mostrou maior estabilidade entre faixas diferentes de complexidade textual, enquanto o GPT-4o-mini demonstrou sensibilidade elevada a respostas com TTR intermediário. Esse comportamento reforça que a precisão do ASAG não depende apenas do modelo ou do *prompt*, mas também da estrutura linguística da resposta analisada.

Referências

ABONIZIO, H. et al. *Sabiá-3 Technical Report*. 2025. Disponível em: <https://arxiv.org/abs/2410.12049>. Citado na página 15.

ABONIZIO, H. et al. *Sabiá-3 Technical Report*. 2025. Preprint. Disponível em: <https://arxiv.org/abs/2410.12049>. Citado na página 10.

AHMED, A.; JOORABCHI, A.; HAYES, M. J. On deep learning approaches to automated assessment: Strategies for short answer grading. *CSEDU (2)*, p. 85–94, 2022. Citado na página 7.

BANG, Y. et al. *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity*. 2023. Disponível em: <https://arxiv.org/abs/2302.04023>. Citado 2 vezes nas páginas 2 e 6.

BISHOP, C. *Pattern Recognition and Machine Learning*. [S.l.: s.n.], 2006. v. 16. 140-155 p. Citado na página 3.

BONTHU, S.; SREE, S. R.; PRASAD, M. H. M. K. Automated short answer grading using deep learning: A survey. In: *Machine Learning and Knowledge Extraction*. Berlin, Heidelberg: Springer-Verlag, 2021. p. 61–78. Citado na página 1.

BONTHU, S.; SRIPADA, R. S.; PRASAD, M. Automated short answer grading using deep learning: A survey. In: _____. [S.l.: s.n.], 2021. p. 61–78. ISBN 978-3-030-84059-4. Citado 2 vezes nas páginas 4 e 5.

BROWN, T. B. et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020. v. 33, p. 1877–1901. Disponível em: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf>. Citado na página 10.

BURROWS, S.; GUREVYCH, I.; STEIN, B. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, v. 25, p. 60–117, 2015. Citado 3 vezes nas páginas 1, 7 e 8.

CAMUS, L.; FILIGHERA, A. 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*. Berlin, Heidelberg: Springer-Verlag, 2020. p. 43–48. ISBN 978-3-030-52239-1. Disponível em: <https://doi.org/10.1007/978-3-030-52240-7_8>. Citado na página 7.

CARPENTER, D. et al. 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)*. Mexico City, Mexico: Association for Computational Linguistics, 2024. p. 403–413. Citado 4 vezes nas páginas 2, 7, 15 e 20.

CHAMIEH, I.; ZESCH, T.; GIEBERMANN, K. LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In: KOCHMAR, E. et al. (Ed.). *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational*

Applications (BEA 2024). Mexico City, Mexico: Association for Computational Linguistics, 2024. p. 309–315. Disponível em: <https://aclanthology.org/2024.bea-1.25/>. Citado na página 2.

CHANG, L.-H.; GINTER, F. Automatic short answer grading for finnish with chatgpt. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 38, n. 21, p. 23173–23181, Mar. 2024. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/30363>. Citado 3 vezes nas páginas 6, 7 e 8.

DEVLIN, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. Disponível em: <https://arxiv.org/abs/1810.04805>. Citado na página 3.

ELNAKA, A. et al. Arascore: Investigating response-based arabic short answer scoring. *Procedia Computer Science*, v. 189, p. 282–291, 2021. ISSN 1877-0509. AI in Computational Linguistics. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050921012114>. Citado na página 1.

FLEISS, J. L.; COHEN, J.; EVERITT, B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, v. 72, n. 5, p. 323–327, 1969. Disponível em: <https://doi.org/10.1037/h0028106>. Citado na página 11.

FREITAG, R. M. K.; GOIS, T. S. d. 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)*. Sociedade Brasileira de Computação, 2024. (STIL 2024), p. 317–326. Disponível em: <http://dx.doi.org/10.5753/stil.2024.241891>. Citado na página 7.

GALHARDI, L. et al. Portuguese automatic short answer grading. In: . [S.l.: s.n.], 2018. p. 1373. Citado na página 8.

GALHARDI, L.; SOUZA, R. de; BRANCHER, J. Automatic grading of portuguese short answers using a machine learning approach. In: *Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação*. [S.l.: s.n.], 2020. p. 109–124. Citado 3 vezes nas páginas 8, 9 e 15.

GIRAY, L. Prompt engineering with chatgpt: A guide for academic writers. *Annals of Biomedical Engineering*, v. 51, n. 12, p. 2629–2633, 2023. Disponível em: <https://doi.org/10.1007/s10439-023-03272-4>. Citado na página 10.

GRéVISSE, C. Llm-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, v. 24, n. 1, p. 1060, 2024. ISSN 1472-6920. Disponível em: <https://doi.org/10.1186/s12909-024-06026-5>. Citado na página 2.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. ed. Springer Series in Statistics, 2009. ISBN 978-0387848570. Disponível em: <https://link.springer.com/book/10.1007/978-0-387-84858-7>. Citado na página 3.

IMRAN, M.; ALMUSHARRAF, N. Google gemini as a next generation ai educational tool: A review of emerging educational technology. *Smart Learning Environments*, v. 11, n. 1, p. 22, 2024. ISSN 2196-7091. Disponível em: <https://doi.org/10.1186/s40561-024-00310-z>. Citado na página 10.

KHOT, T. et al. *Decomposed Prompting: A Modular Approach for Solving Complex Tasks*. 2023. Preprint. Disponível em: <https://arxiv.org/abs/2210.02406>. Citado 3 vezes nas páginas 6, 7 e 10.

LIU, P. et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 55, n. 9, jan. 2023. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3560815>. Citado na página 6.

LIU, Q.; KUSNER, M. J.; BLUNSOM, P. *A Survey on Contextual Embeddings*. 2020. Disponível em: <https://arxiv.org/abs/2003.07278>. Citado na página 1.

MALEKZADEH, M. et al. Review of graph neural network in text classification. In: *2021 IEEE 12th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*. [S.l.: s.n.], 2021. p. 0084–0091. Citado na página 5.

MANNING, C. D.; SCHüTZE, H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999. ISBN 978-0262133609. Citado na página 4.

MELLO, R. et al. Prompt engineering for automatic short answer grading in brazilian portuguese. In: *Anais do XXXV Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2024. p. 1730–1743. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/31353>. Citado 6 vezes nas páginas 2, 6, 7, 8, 9 e 15.

MELLO, R. F. et al. 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)*. New York, NY, USA: Association for Computing Machinery, 2025. p. 93–103. Citado 9 vezes nas páginas 1, 2, 5, 6, 7, 9, 10, 15 e 20.

MITCHELL, T. *Machine Learning*. McGraw-Hill, 1997. (McGraw-Hill International Editions). ISBN 9780071154673. Disponível em: <https://books.google.com.br/books?id=EoYBngEACAAJ>. Citado na página 3.

MOHLER, M.; MIHALCEA, R. Text-to-text semantic similarity for automatic short answer grading. In: LASCARIDES, A.; GARDENT, C.; NIVRE, J. (Ed.). *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece: Association for Computational Linguistics, 2009. p. 567–575. Disponível em: <https://aclanthology.org/E09-1065/>. Citado na página 7.

MOTA, C. et al. Classificação de páginas de petições iniciais utilizando redes neurais convolucionais multimodais. In: *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*. Porto Alegre, RS, Brasil: SBC, 2020. p. 318–329. ISSN 2763-9061. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/12139>. Citado na página 5.

NOVAK, B.; MLADENIČ, D.; GROBELNIK, M. Text classification with active learning. In: SPILIOPOULOU, M. et al. (Ed.). *From Data and Information Analysis to Knowledge Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 398–405. ISBN 978-3-540-31314-4. Citado na página 4.

OLSSON, F. *A literature survey of active machine learning in the context of natural language processing*. 1. ed. [S.l.], 2009. 59 p. (SICS Technical Report, 2009:06). Citado na página 3.

PERUMAL, T. et al. A comprehensive overview and comparative analysis on deep learning models. *Journal on Artificial Intelligence*, Tech Science Press, v. 6, n. 1, p. 301–360, 2024. ISSN 2579-003X. Disponível em: <http://dx.doi.org/10.32604/jai.2024.054314>. Citado na página 5.

PIECH, C. et al. Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*, 2013. Citado na página 1.

QIN, L. et al. *Large Language Models Meet NLP: A Survey*. 2024. Preprint. Disponível em: <https://arxiv.org/abs/2405.12819>. Citado na página 2.

RAWTE, V.; SHETH, A.; DAS, A. *A Survey of Hallucination in Large Foundation Models*. 2023. Disponível em: <https://arxiv.org/abs/2309.05922>. Citado na página 12.

RIPMIATIN, E.; PURNAMASARI, P. D.; RATNA, A. A. P. 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*. New York, NY, USA: Association for Computing Machinery, 2024. (ICCIP '23), p. 492–497. ISBN 9798400708909. Disponível em: <https://doi.org/10.1145/3638884.3638962>. Citado na página 7.

SANTU, S. K. K.; FENG, D. TELeR: A general taxonomy of LLM prompts for benchmarking complex tasks. In: BOUAMOR, H.; PINO, J.; BALI, K. (Ed.). *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023. p. 14197–14203. Disponível em: <https://aclanthology.org/2023.findings-emnlp.946/>. Citado na página 6.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM Computing Surveys*, Association for Computing Machinery (ACM), v. 34, n. 1, p. 1–47, mar. 2002. ISSN 1557-7341. Disponível em: <http://dx.doi.org/10.1145/505282.505283>. Citado na página 5.

SETTLES, B. *Active Learning Literature Survey*. [S.l.], 2009. Disponível em: <http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf>. Citado na página 4.

SUNG, C.; DHAMECHA, T. I.; MUKHI, N. 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*. Berlin, Heidelberg: Springer-Verlag, 2019. p. 469–481. ISBN 978-3-030-23203-0. Disponível em: <https://doi.org/10.1007/978-3-030-23204-7_39>. Citado na página 7.

SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. 2. ed. MIT Press, 2018. Disponível em: <http://incompleteideas.net/book/the-book-2nd.html>. Citado na página 3.

SüZEN, N. et al. Automatic short answer grading and feedback using text mining methods. In: *Procedia Computer Science*. [S.l.: s.n.], 2020. v. 169, p. 726–743. Citado na página 1.

VASWANI, A. et al. *Attention Is All You Need*. 2023. Disponível em: <https://arxiv.org/abs/1706.03762>. Citado na página 6.

WARNEKE, K. et al. The impact of sample size on reliability metrics stability in isokinetic strength assessments: Does size matter? *Measurement in Physical Education and Exercise Science*, Routledge, v. 0, n. 0, p. 1–12, 2025. Disponível em: <https://doi.org/10.1080/1091367X.2025.2494998>. Citado na página 12.

WEI, J. et al. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. Citado na página 2.

WEI, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022. v. 35, p. 24824–24837. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf>. Citado 2 vezes nas páginas 7 e 10.

WHITE, J. et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023. Disponível em: <https://arxiv.org/abs/2302.11382>. Citado na página 10.

XU, Z.; JAIN, S.; KANKANHALLI, M. *Hallucination is Inevitable: An Innate Limitation of Large Language Models*. 2025. Disponível em: <https://arxiv.org/abs/2401.11817>. Citado 3 vezes nas páginas 6, 12 e 19.

YAN, L. et al. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, v. 55, p. 90–112, 2024. Citado 2 vezes nas páginas 2 e 6.

ZHAO, W. X. et al. *A Survey of Large Language Models*. 2025. Disponível em: <https://arxiv.org/abs/2303.18223>. Citado na página 6.

ZHAO, Z. et al. Improved llm methods using linear regression. In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Fort Worth, TX, USA: IEEE, 2017. p. 5350–5353. Citado na página 11.

ZHUANG, Z. et al. *Through the Lens of Core Competency: Survey on Evaluation of Large Language Models*. 2023. Disponível em: <https://arxiv.org/abs/2308.07902>. Citado na página 1.