



FEDERAL DO ESTADO DO RIO DE UNIVERSIDADE
JANEIRO CENTRO DE CIÊNCIAS EXATAS E
TECNOLOGIA PROGRAMA DE PÓS-GRADUAÇÃO EM
INFORMÁTICA

Estudo sobre o uso de Meta-Prompting em LLMs para a Elaboração de Exercícios:
Na Perspectiva de Docentes

Wendy Silva Cataldo

Orientador

Dr. Carlos Eduardo De Mello

Coorientadora

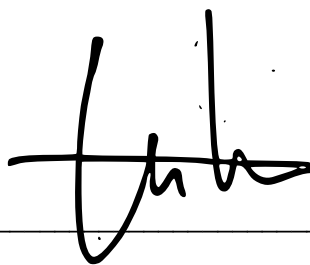
Dra. Laura de Oliveira Fernandes Moraes

Estudo sobre o uso de Meta-prompting em LLMs para a Elaboração de Exercícios: Na
Perspectiva de Docentes

Wendy Silva Cataldo

PROPOSTA DE DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA
OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM
INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
(UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:



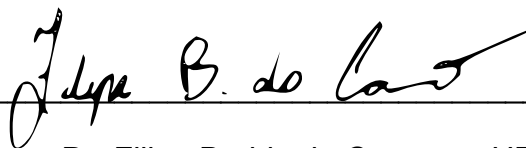
Dr. Carlos Eduardo Ribeiro de Mello — UNIRIO



Dra. Laura de Oliveira Fernandes Moraes — UNIRIO



Dr. Sean Wolfgang Matsui Siqueira — UNIRIO



Dr. Filipe Braida do Carmo — UFRRJ

Catálogo informatizada pelo(a) autor(a)

S357 Silva Cataldo, Wendy
 Estudo sobre o uso de Meta-Prompting em LLMs para a
Elaboração de Exercícios: Na Perspectiva de Docentes / Wendy
Silva Cataldo. -- Rio de Janeiro : UNIRIO, 2025.
 161

 Orientador: Carlos Eduardo Ribeiro de Mello .
 Coorientadora: Laura de Oliveira Fernandes Moraes .
 Dissertação (Mestrado) - Universidade Federal do Estado
do Rio de Janeiro, Programa de Pós-Graduação em Informática,
2025.

 1. Meta-prompt. 2. Modelos de Linguagem de Larga
Escala. 3. Geração de Perguntas Educacionais. I. Mello ,
Carlos Eduardo Ribeiro de, orient. II. Moraes , Laura de
Oliveira Fernandes , coorient. III. Título.

Dedico este trabalho à memória de meu pai, Roberto Luiz Pereira, cuja presença em meu coração me acompanhou em cada etapa desta jornada. À minha avó Neyde Saraiva Pereira, minha eterna fonte de incentivo. Dedico este trabalho como um gesto de gratidão por todo o amor e apoio incondicionais.

AGRADECIMENTOS

Agradeço a Deus por me proporcionar esta experiência única e enriquecedora, que me permitiu crescer pessoal e profissionalmente.

Agradeço às mulheres da minha família, por serem meus maiores exemplos de força e determinação. Seu amor, apoio incondicional e crença em meu potencial foram a base para todas as minhas conquistas pessoais. Agradeço ao meu namorado, por sua paciência, companheirismo e amor incondicional, que foram fundamentais para superar os desafios desta jornada.

Agradeço aos meus amigos pelo apoio constante, pela amizade sincera e por todos os momentos de alegria que compartilhamos durante essa jornada.

Quero agradecer também ao meu orientador Carlos Eduardo e a coorientadora Laura de Oliveira por me orientar.

Esta dissertação é fruto de muita dedicação, trabalho duro e do sacrifício de momentos preciosos ao lado de pessoas queridas.

Com todo meu carinho e gratidão.

Cataldo, Wendy Silva **Estudo sobre o uso de Meta-Prompting em LLMs para a Elaboração de Exercícios: Na Perspectiva de Docentes**. UNIRIO, 2025. 161 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

Resumo

Esta dissertação investiga o uso da técnica de *Meta-Prompt* em *Large Language Models* (LLMs) para a geração automatizada de exercícios educacionais, a partir da perspectiva de docentes da disciplina de Língua Portuguesa. O objetivo central foi avaliar se a aplicação de *Meta-Prompt* eleva a qualidade das questões produzidas por LLMs, em comparação com *Prompts* convencionais. O estudo adota uma abordagem experimental e comparativa, na qual questões geradas por LLMs com e sem o uso de *Meta-Prompt* foram analisadas em relação às elaboradas por professores humanos. Foram utilizados critérios de avaliação pedagógica amplamente reconhecidos: cobertura, gramaticalidade, relevância, utilidade e capacidade de resposta, complementados por análise estatística descritiva e inferencial. Os dados foram coletados por meio de questionários aplicados a docentes, que avaliaram a qualidade das questões geradas. Os resultados indicam que, embora a aplicação do *Meta-Prompt* tenha promovido ganhos qualitativos na clareza e na contextualização das questões, as diferenças observadas não foram estatisticamente significativas em todos os critérios. Contudo, verificou-se maior consistência linguística e aderência aos objetivos educacionais quando o *Meta-Prompt* foi utilizado. Além disso, a percepção dos professores revelou uma tendência positiva quanto ao potencial pedagógico dos LLMs como ferramentas de apoio à elaboração de materiais didáticos, ressaltando a importância de seu uso ético e supervisionado. A pesquisa contribui para o campo da Inteligência Artificial na Educação ao oferecer evidências empíricas sobre a eficácia do *Meta-Prompt* e ao propor diretrizes para seu emprego em contextos educacionais reais.

Palavras-chave: *Meta-prompt*, Modelos de Linguagem de Larga Escala, Geração de Perguntas Educacionais, Avaliação educacional

ABSTRACT

This dissertation investigates the use of the Meta-Prompt technique in Large Language Models (LLMs) for the automated generation of educational exercises, from the perspective of teachers of the Portuguese language. The main objective was to evaluate whether the application of Meta-Prompt improves the quality of questions produced by LLMs compared to conventional prompts. The study adopts an experimental and comparative approach, in which questions generated by LLMs with and without the use of Meta-Prompt were analyzed in relation to those developed by human teachers. Widely recognized pedagogical evaluation criteria were employed—coverage, grammaticality, relevance, usefulness, and answerability—complemented by descriptive and inferential statistical analysis. Data were collected through questionnaires administered to teachers who assessed the quality of the generated questions. The results indicate that, although the application of the Meta-Prompt promoted qualitative improvements in the clarity and contextualization of the questions, the observed differences were not statistically significant across all criteria. However, greater linguistic consistency and adherence to educational objectives were observed when the Meta-Prompt was applied. Furthermore, teachers' perceptions revealed a positive tendency regarding the pedagogical potential of LLMs as tools to support the development of instructional materials, emphasizing the importance of their ethical and supervised use. This research contributes to the field of Artificial Intelligence in Education by providing empirical evidence of the effectiveness of Meta-Prompting and by proposing guidelines for its application in real educational contexts.

Keywords: Meta-prompt, Large Language Models (LLMs), Educational Question Generation, Educational Assessment.

Sumário

1	INTRODUÇÃO	13
1.1	Contexto e Motivação	13
1.2	Justificativa e relevância	15
1.3	Hipótese	16
1.4	Questões de Pesquisa	16
1.5	Objetivos	17
1.5.1	Objetivo primário	17
1.5.2	Objetivo secundário	17
1.6	Contribuições Alcançadas	17
1.7	Estrutura da dissertação	18
2	FUNDAMENTAÇÃO TÉCNICA	19
2.1	Educação	19
2.2	Definição de Questões e Avaliação no Contexto Educacional	20
2.3	Base Nacional Comum Curricular (BNCC)	21
2.4	Taxonomia de Bloom Revisada	22
2.5	CrITÉRIOS de avaliação das questões	24
2.6	Inteligência Artificial (IA)	27
2.7	Processamento de Linguagem Natural (PLN)	28
2.8	Modelos de Linguagem de Larga Escala (LLM)	29
2.9	Engenharia de Prompts	31
2.9.1	Conceito e tipos de Engenharia de Prompt	31
2.9.2	<i>Meta – Prompt</i>	33
3	REVISÃO DA LITERATURA	35
3.1	Planejamento	35

3.2	Investigação do problema.....	35
3.3	Estudos de Referência	38
4	METODOLOGIA.....	43
4.1	Seleção e teste dos LLMs apropriados.....	44
4.2	Refinamento dos <i>Prompts</i> para geração das questões.....	52
4.3	Avaliação das questões geradas.....	54
5	RESULTADOS E DISCUSSÃO.....	61
5.1	Resultados.....	61
5.1.1	Análise Automática.....	61
5.1.2	Análise do perfil dos respondentes.....	64
5.1.3	Análise da avaliação do Questionário.....	68
5.2	Discussão	73
5.3	Confirmação/Refutação das Hipóteses	75
5.4	Resposta às Questões de Pesquisa.....	76
6	CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS	78
6.1	Limitações do Estudo	78
6.2	Trabalhos Futuros	79
	REFERÊNCIA.....	80
	Apêndice A – Questões criadas e <i>Prompts</i> usados para os testes preliminares	88
	Apêndice B – Questões criadas e Prompts usados para o experimento	123
	Apêndice C – Termo de Consentimento Livre e Esclarecido (TCLE) e Questionário	138
	Apêndice D – E-mail para convidar os Professores.....	146
	Apêndice E - Resultados Detalhados	148
	Apêndice F – Manual para os professores	158

LISTA DE FIGURAS

Figura 1 - Comparação entre a taxonomia de Bloom original e a taxonomia revisada.	23
Figura 2 - Estrutura hierárquica da inteligência artificial, evidenciando a progressão de conceitos de ia geral até a especialização em modelos generativos.	28
Figura 3 - Evolutiva dos modelos de linguagem de larga escala.....	31
Figura 4 – Processo de avaliação de LLMs.....	44
Figura 5 - Fluxograma do processo analítico quantitativo.	59

LISTA DE QUADRO

Quadro 1 - Prompt e <i>Meta-Prompt</i> usados nos testes preliminares.....	48
Quadro 2 - Meta-Prompt revisado para o teste.	52
Quadro 3 - Prompt, Meta-Prompt e pedido para a professora para criarem as questões para o experimento.	53
Quadro 4 - Criação do Prompt para que os LLMs avaliem as questões.	55

LISTA DE TABELA

Tabela 1 - Critérios de inclusão e exclusão utilizados.....	35
Tabela 2 - Comparação entre os LLMS.....	45
Tabela 3 - Análise cruzada com avaliadores LLMS agregados na escala 1–5, reportando Mediana, Q1 (25%), Q3 (75%) e n por Critério × Gerador × Tipo de Prompt; LLMS avaliados em Prompt e Meta-prompt, Professor apenas em Prompt.	62
Tabela 4 - Perfil dos professores: distribuição de respostas por variável (frequências e percentuais).....	65
Tabela 5 - Bem-estar docente (variáveis ordinais): estresse e satisfação na escala Likert 1 – 5.....	66
Tabela 6 — Estatísticas globais por critério (n, mediana, Q1 e Q3).....	68
Tabela 7 — Estatísticas por critério × gerador × tipo (n, mediana, Q1 e Q3).	69

LISTA DE NOMENCLATURAS E SIGLAS

AAAI	<i>Association for the Advancement of Artificial Intelligence</i>
BNCC	Base Nacional Comum Curricular
DL	<i>Deep Learning</i> (Aprendizado Profundo)
GPE	Geração de Perguntas Educacionais - EQG, do Inglês <i>Educational Question Generation</i>
IA	Inteligência Artificial
IC95%	Intervalo de Confiança 95%
IQR	Intervalo Interquartil
LDB	Lei de Diretrizes e Bases da Educação
LGPD	Lei geral de proteção de dados
LLM	<i>Large Language Model</i> (Modelo de Linguagem Grande)
MEC	Ministério da Educação
ML	<i>Machine Learning</i> (Aprendizado de Máquina)
PLN	Processamento de Linguagem Natural
PNE	Plano Nacional de Educação
Saeb	Sistema de Avaliação da Educação Básica
TCLE	Termo de Consentimento Livre e Esclarecido

1.1 Contexto e Motivação

A educação, pilar fundamental para o desenvolvimento social, tem nos professores seus principais agentes. Nesse contexto, a Inteligência Artificial (IA) apresenta-se como ferramenta promissora para otimizar o trabalho dos docentes, com potencial para auxiliar na criação de exercícios (Cardoso *et al.*, 2023; Santos *et al.*, 2024). A IA abrange diversas aplicações, desde a criação de materiais didáticos personalizados até a automação de processos administrativos, liberando os professores para se dedicarem ao ensino e ao atendimento individualizado dos alunos (Santos *et al.*, 2024). A customização do conteúdo, com informações geradas por ferramentas de IA, pode atender às necessidades de cada professor, tornando o ensino mais relevante (Picão *et al.*, 2023).

Contudo, a elaboração de atividades que atendam às necessidades pedagógicas e promovam um aprendizado ativo é trabalhosa para boa parte dos educadores. Diante deste cenário, pretende-se analisar como os professores da disciplina de Língua Portuguesa podem se beneficiar da utilização da *Large Language Model* (LLM) (Santos *et al.*, 2024). As aplicações da IA na educação variam desde assistentes virtuais, que fornecem suporte sem intervenção humana, até sistemas de avaliação automatizados que auxiliam na elaboração de exames. Estudar o uso dos LLMs na educação é uma forma de buscar soluções que possam agregar valor para o processo de ensino e aprendizagem, apoiando os professores. Desenvolver um conjunto abrangente de perguntas em um contexto específico é essencial para os educadores, uma vez que possibilita uma exploração e avaliação detalhadas da compreensão dos alunos (Al Faraby; Adiwijaya; Romadhony, 2024). Essa abordagem assegura que os educadores abranjam uma variedade ampla de objetivos de aprendizagem e níveis cognitivos, facilitando assim um exame mais aprofundado do conteúdo. Adicionalmente, a elaboração de um conjunto de perguntas permite que os educadores considerem diferentes estilos e preferências de aprendizagem dos alunos, promovendo a inclusão e melhorando os resultados educacionais de forma significativa. De modo geral, a capacidade de criar um conjunto de perguntas contextualizadas habilita os educadores a desenhar

experiências de aprendizagem enriquecedoras, que suportam o desenvolvimento integral dos alunos e a aquisição plena do conhecimento.

Nos últimos anos, o avanço dos LLMs vem revolucionando a educação. Modelos, como o GPT-4 (*Generative Pretrained Transformer*) e o *Mistral*, têm demonstrado grande potencial na automatização de tarefas educacionais, como a Geração de Perguntas Educacionais (GPE) e respostas para avaliações (Chan; An; Davoudi, 2023; Ionescu; Enescu, 2023). A capacidade desses modelos em gerar conteúdos relevantes tem despertado o interesse de educadores, que buscam explorar seu potencial para apoiar a elaboração de exercícios e melhorar a experiência de aprendizagem (Ionescu; Enescu, 2023). Estudos recentes têm investigado a eficácia dos LLMs na GPE e respostas para diferentes contextos educacionais. Por exemplo, (Chan; An; Davoudi, 2023) demonstraram que o *ChatGPT* pode competir com modelos de estado da arte na GPE, especialmente quando utilizado com *prompts* bem estruturados. Além disso, a utilização de plataformas baseadas em IA para a geração de *quizzes* tem sido explorada como uma forma de reduzir a carga de trabalho dos educadores e personalizar a aprendizagem dos alunos (Ionescu; Enescu, 2023). No entanto, os desafios na avaliação de respostas complexas e a necessidade de supervisão humana permanecem como pontos críticos a serem superados (Chan; An; Davoudi, 2023).

O Processamento de Linguagem Natural (PLN) foca na análise e compreensão da linguagem natural e seu uso em várias aplicações, como a elaboração de questões. A GPE consiste em formular perguntas que sejam relevantes e coerentes dentro de um contexto definido, tendo como referência textos escritos em linguagem natural, sendo necessário entender o contexto para fornecer as respostas.

A customização do conteúdo educacional, com base em dados provenientes da IA, pode satisfazer às necessidades de cada professor, tornando o ensino mais relevante (Picão *et al.*, 2023). Globalmente, a utilização da IA no âmbito da educação tem sido amplamente explorada e pesquisada, transformando o cenário educacional. Em diversas partes do mundo, instituições de ensino estão explorando essas possibilidades tecnológicas e muitas estão integrando a IA às suas práticas pedagógicas (Oliveira *et al.*, 2023). A IA consiste em um conjunto de

tecnologias que viabilizam a criação de sistemas com capacidade para realizar tarefas previamente restritas aos seres humanos, como reconhecimento de imagens e tomada de decisões (Gomes, 2010).

No entanto, o uso da IA na educação também traz desafios éticos e responsáveis. Questões como a privacidade dos dados dos alunos, a possibilidade de viés nos algoritmos e a dependência excessiva da tecnologia precisam ser cuidadosamente consideradas (Gomes, 2010). Além disso, é fundamental garantir que a IA seja utilizada de forma complementar ao trabalho dos professores, e não como um substituto, para evitar a perda da dimensão humana no processo de ensino e aprendizagem (Oliveira *et al.*, 2023). A criação de exercícios que atendam às necessidades pedagógicas e promovam o aprendizado ativo é desafiadora para boa parte dos educadores (Tran *et al.*, 2023; Wang *et al.*, 2022). Diante deste cenário, pretende-se discutir como os professores de disciplinas relacionadas à Língua Portuguesa podem se beneficiar da utilização do LLMs, garantindo que os benefícios dessa tecnologia sejam aproveitados de forma responsável e ética (Santos *et al.*, 2024).

Diante deste cenário, o presente projeto visa avaliar as respostas de professores de Língua Portuguesa sobre a criação de questões com o uso de LLMs. O estudo busca entender como a tecnologia pode otimizar e melhorar a experiência na criação de materiais didáticos. Além disso, o projeto pretende contribuir para a discussão sobre o papel da IA na educação, explorando suas potencialidades e limitações, bem como destacando a importância do uso responsável e ético dessa tecnologia.

1.2 Justificativa e relevância

A integração da IA e dos LLMs demonstra grande potencial na educação, especialmente na GPE. Essa capacidade é reforçada pela necessidade de otimizar o trabalho do docente e aprimorar a qualidade do ensino, oferecendo uma oportunidade para auxiliar os professores na elaboração de diversos tipos de perguntas.

A automação na GPE não só economiza tempo para os educadores, mas também permite que eles se concentrem no desenvolvimento de estratégias

pedagógicas inovadoras. Além disso, a personalização possibilitada por essa tecnologia pode levar a experiências de aprendizado mais eficazes. À medida que os LLMs continuam a evoluir, sua integração às atividades de ensino torna-se crucial para maximizar seu potencial pedagógico, oferecendo um suporte que pode revolucionar o cenário do ensino.

Técnicas de *Meta-Prompt* permitem orientar os LLMs para a geração de perguntas mais claras, coerentes, contextualizadas e alinhadas aos objetivos curriculares. Ao focar na disciplina de Língua Portuguesa, o estudo também contribui para suprir uma lacuna da literatura, que tende a concentrar-se em áreas como Ciências Exatas e Programação. Dessa forma, a presente pesquisa se justifica pela sua originalidade, relevância prática para os professores, e pela contribuição potencial para o avanço do uso ético e eficiente da IA no contexto educacional.

1.3 Hipótese

A hipótese deste trabalho é que a utilização de técnicas como o *Meta-Prompt* em LLMs melhora a qualidade das perguntas geradas, em comparação com aquelas elaboradas sem tal técnica.

- **H1:** A aplicação de *Meta-Prompt* em LLMs eleva a qualidade das questões geradas em comparação com o uso de *Prompts*, refletindo-se em uma pontuação superior nos critérios definidos.
- **H2:** LLMs, quando adequadamente instruídos, são capazes de gerar questões de Língua Portuguesa com qualidade pedagógica equivalente à de um professor, não havendo diferença estatisticamente significativa entre as avaliações de ambos os grupos.

1.4 Questões de Pesquisa

Para esta dissertação, foram definidas as seguintes questões de pesquisa (QP):

QP1: A utilização de *Meta-Prompt* em LLMs resulta em questões de Língua Portuguesa com qualidade pedagógica significativamente superior às aquelas

geradas por meio de prompts simples, na percepção de docentes?

QP2: Questões de Língua Portuguesa geradas por LLMs, com ou sem *Meta-Prompt*, possuem qualidade comparável àquelas elaboradas por um professor especialista, considerando critérios como relevância, utilidade e adequação curricular?

1.5 Objetivos

1.5.1 Objetivo primário

Entender se o uso de LLMs com a utilização de técnicas de *Meta-Prompt* para criar exercícios melhora a qualidade na elaboração das questões. Para isso, iremos comparar por meio de questionários, questões criadas por um professor, por um *Prompt* e com o uso de *Meta-Prompt*. Buscamos, ainda, identificar as vantagens da utilização do *Meta-Prompt* na criação de exercícios. Coletamos dados para verificar a possibilidade de

- Gerar exercícios personalizados, contextualizados e alinhados ao conteúdo programático, utilizando técnicas de *Meta-Prompt* para garantir relevância e adequação;
- Proporcionar aos alunos uma experiência de aprendizagem mais rica e desafiadora, aumentando a variedade e a qualidade das questões.

1.5.2 Objetivo secundário

O objetivo secundário é avaliar a eficácia de LLMs ao empregar técnicas de *Meta-Prompt* na GPE e contextualizados para professores de Língua Portuguesa. Esta avaliação se concentra na avaliação da qualidade e relevância desses exercícios.

1.6 Contribuições Alcançadas

Esta dissertação contribui para o avanço científico ao investigar o impacto do *Meta-Prompt* na qualidade das questões geradas por *Large Language Models* (LLMs) para a disciplina de Língua Portuguesa. A pesquisa oferece uma análise comparativa entre as questões geradas por LLMs com e sem *Meta-Prompt* e as

questões elaboradas por um professor especialista, fornecendo *insights* valiosos sobre o potencial e as limitações dessas tecnologias no contexto educacional. Além disso, a análise da influência do perfil docente nas avaliações contribui para a compreensão dos fatores que modulam a aceitação e o uso de LLMs na educação.

1.7 Estrutura da dissertação

Esta pesquisa está organizada da seguinte forma: no Capítulo 1 foram abordados pontos chaves que orientam toda a dissertação, começando pela Motivação, seguida pelo Contexto, Justificativa, Hipótese, Objetivos, Metodologia e Contribuições Esperadas. No Capítulo 2 é realizada a Fundamentação Teórica das ferramentas utilizadas no experimento. No Capítulo 3 é apresentada a Revisão da Literatura. No Capítulo 4 é apresentada a Metodologia. No Capítulo 5 são apresentados os resultados e uma discussão. No Capítulo 6 são apresentados os comentários, as limitações do estudo e as considerações finais.

2 FUNDAMENTAÇÃO TÉCNICA

Os conceitos discutidos neste capítulo serão utilizados ao longo deste estudo, facilitando a compreensão da abordagem adotada. Da subseção 2.1 a 2.4 serão abordados temas referentes à educação, na subseção 2.5, serão abordados os critérios usados para avaliar as questões criadas e da subseção 2.6 a 2.9 serão abordados os princípios da IA.

2.1 Educação

A educação é um processo dinâmico e contínuo que visa ao desenvolvimento integral do indivíduo, abrangendo aspectos cognitivos, emocionais, sociais e culturais. Para Freire (1996), a educação é um ato de amor, um compromisso com o desenvolvimento humano e a transformação social. Esse processo não se limita à mera transmissão de conhecimentos, mas envolve a formação crítica e reflexiva dos sujeitos, preparando-os para a participação ativa na sociedade (Freire, 1996). Freire (1970) introduz o conceito de educação como prática da liberdade, onde o ensino deve promover a consciência crítica e a autonomia dos alunos. Esse enfoque pedagógico defende que a educação deve ser um meio para que os indivíduos compreendam e transformem sua realidade, engajando-se em um processo de diálogo e reflexão crítica. Assim, a educação assume um papel transformador, não apenas individualmente, mas também socialmente.

Os professores são agentes fundamentais no processo educativo, atuando como mediadores entre o conhecimento e os alunos. Vargas (2008) destaca a importância do saber docente, que vai além do conhecimento disciplinar, incluindo o saber pedagógico e a capacidade de relacionar-se com os alunos de maneira significativa. A formação e o desenvolvimento contínuo dos professores são essenciais para que possam desempenhar efetivamente esse papel.

O cenário atual apresenta diversos desafios para a educação, incluindo a necessidade de adaptação às rápidas mudanças tecnológicas e às demandas de uma sociedade globalizada. Moran (2007) aponta que a educação deve ser flexível e inovadora, incorporando novas tecnologias e metodologias que atendam às

necessidades dos estudantes do século XXI. Isso inclui a integração de ferramentas digitais e a personalização do ensino.

A educação também desempenha um papel crucial na promoção da inclusão social e na redução das desigualdades. De acordo com (Abramovay *et al.*, 2003), uma educação inclusiva deve garantir o acesso e a permanência de todos os estudantes, independentemente de suas condições sociais, econômicas, culturais ou físicas. Isso requer políticas educacionais que promovam a equidade e a justiça social, assegurando que todos os indivíduos tenham a oportunidade de desenvolver seu potencial.

A educação é um processo complexo e multifacetado que envolve não apenas a aquisição de conhecimentos, mas também o desenvolvimento de habilidades críticas, sociais e emocionais. Ela é fundamental para o desenvolvimento humano e social, promovendo a inclusão, a equidade e a transformação. Os desafios contemporâneos exigem uma educação adaptável e inovadora, capaz de preparar os indivíduos para uma sociedade em constante mudança. Com a expansão do acesso às tecnologias de IA, a educação tem a oportunidade de assumir um papel cada vez mais significativo na melhoria e complementação de seus sistemas escolares, que estão sobrecarregados.

2.2 Definição de Questões e Avaliação no Contexto Educacional

Nesta dissertação, empregamos o termo avaliação de forma instrumental refere-se à atribuição que apoiam as questões geradas (por professor e por LLMs), visando julgar mérito, adequação e utilidade pedagógica sob critérios explícitos, mantendo alinhamento entre objetivos de aprendizagem e evidências coletadas.

A avaliação é um processo central para o ensino e a gestão educacional: fornece evidências sobre aprendizagens e efetividade pedagógica, apoiando decisões de professores, pesquisadores e formuladores de políticas (Ewell, 2008; Linn, 2003; Nagy, 2000). A qualidade da avaliação depende, de modo decisivo, da qualidade dos itens: as questões precisam medir o que se propõem a medir, com clareza e coerência entre objetivos, conteúdo e critérios (Kliebard, 2011).

Falhas na elaboração de itens afetam especialmente modelos personalizados e adaptativos de ensino, que exigem acervos robustos e consistentes de perguntas por isso, a qualidade se torna um pilar para práticas inovadoras (Cronbach, 1963; Scriven, 1991; Thelen; Tyler, 1950).

Nesta dissertação, questões educacionais são entendidas como itens que eliciam evidências de aprendizagem ao demandar do estudante operações cognitivas pertinentes ao objetivo de ensino (p.ex., recordar, interpretar, aplicar, analisar). Em sala de aula, tais itens funcionam simultaneamente como instrumentos de avaliação e aprendizagem, promovendo reflexão, análise e aplicação de conhecimentos (Howard; Aleman, 2008).

A qualidade do item depende de formulação clara, foco no construto avaliado e adequação ao nível dos estudantes; itens bem planejados não apenas verificam conhecimento, mas também induzem exploração conceitual e pensamento de ordem superior (Billing, 2007). Assim, adotamos uma definição operacional: cada questão específica contexto, habilidade/nível cognitivo, resposta-alvo e critérios de julgamento usados nas análises desta pesquisa. Na prática, consideramos questões item usuais (objetivos e abertos), desde que atendam aos requisitos anteriores e ao alinhamento curricular discutido nas próprias subseções (BNCC e Taxonomia Bloom). Essa delimitação conecta a definição conceitual ao uso que efetivamente fazemos das questões no experimento (geração por LLMs e por docente, seguida de avaliação humana e automática).

2.3 Base Nacional Comum Curricular (BNCC)

A Base Nacional Comum Curricular (BNCC)¹ é um documento normativo e obrigatório do Ministério da Educação (MEC)² que define os direitos de aprendizagem e o conjunto orgânico e progressivo de competências e habilidades para toda a Educação Básica, visando reduzir desigualdades e orientar currículos com base em parâmetros comuns (BRASIL, 2018). Além de equidade educacional, o texto enfatiza formação integral e alinhamento legal Lei de Diretrizes e Bases da

¹ <https://basenacionalcomum.mec.gov.br/>

² <https://www.gov.br/mec/pt-br>

Educação (LDB)³ e Plano Nacional de Educação (PNE)⁴, em articulação com políticas como o Sistema de Avaliação da Educação Básica (Saeb)⁵.

No 7º ano de Língua Portuguesa, a BNCC organiza o ensino em eixos como práticas de leitura/escuta, produção textual, análise linguística e semiótica e oralidade, com objetivos que integram letramentos contemporâneos (argumentação, revisão colaborativa, adequação à norma, respeito à variação linguística). Esses eixos funcionam, nesta pesquisa, como balizas curriculares para definir conteúdos-alvo e habilidades que ancoram os critérios de relevância e cobertura das questões.

Operacionalmente, mapeamos as habilidades da BNCC às tarefas de geração de itens e às rubricas de avaliação. Exemplos frequentes no corpus da pesquisa incluem EF69LP03 (inferência lexical em contexto), EF07LP20 (recursos verbais/não verbais em textos multissemióticos) e EF69LP11 (produção de narrativas), utilizados para delimitar o escopo das questões e verificar alinhamento curricular.

2.4 Taxonomia de Bloom Revisada

Adotamos a Taxonomia de *Bloom* revisada como referência para classificar a exigência cognitiva das questões geradas e avaliadas neste estudo. A revisão de Anderson e Krathwohl (2001) atualiza a terminologia (substitui “Síntese” por “Criação”) e reordena as categorias, posicionando Criar no nível mais alto; seu propósito é oferecer uma estrutura para formular objetivos, atividades e avaliações coerentes ver Figura 1. Essa taxonomia sustenta o desenho e a análise das questões, alinhando níveis cognitivos aos objetivos de aprendizagem e ao julgamento de qualidade.

3

https://www.geledes.org.br/ldb/?gad_source=1&gad_campaignid=1495757196&gbraid=0AAAAADnS6iAXXWRCXiFZzP2gxPu_sP-2r&gclid=CjwKCAjwvO7CBhAqEiwA9q2YJRuKx8DT-S42JUa-YJS0XxizG5PRItNnYJ2LVZc-lqjN4AA5cc2tIxoCCUQQAvD_BwE

⁴ <https://pne.mec.gov.br/>

⁵ <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb>

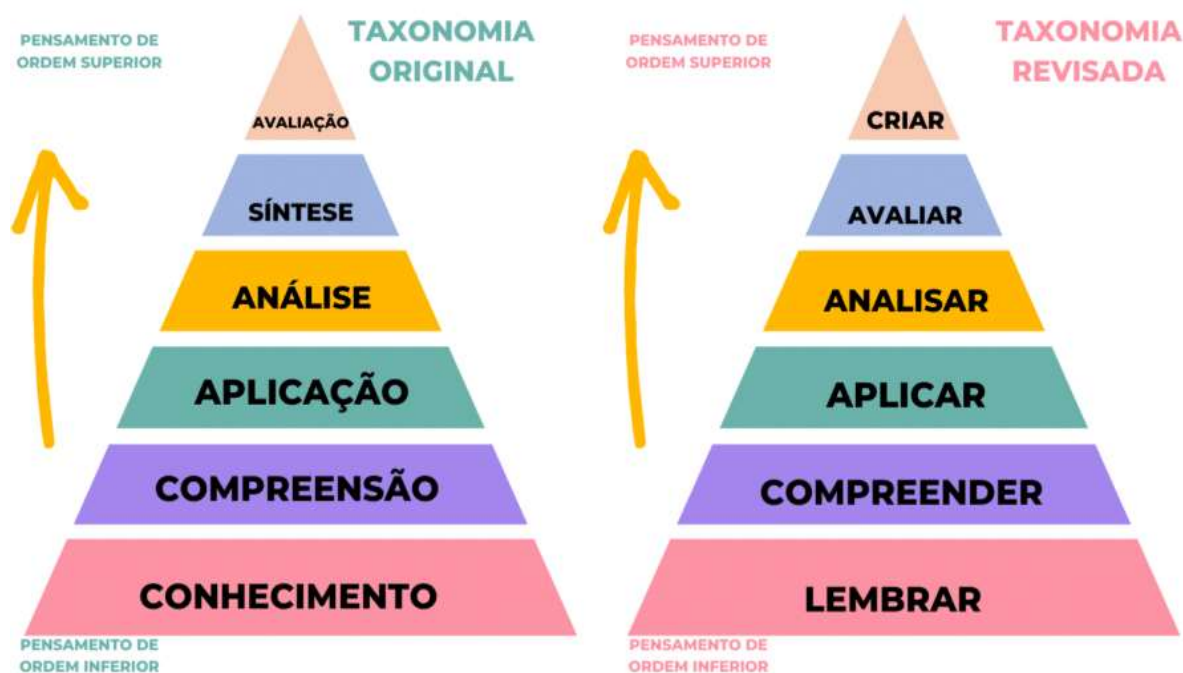


Figura 1 - Comparação entre a taxonomia de Bloom original e a taxonomia revisada.

Fonte: Educador, 2024.

Trabalhamos, portanto, com os seis níveis hierárquicos (do mais básico ao mais complexo), empregados como rótulos operacionais nas análises:

- Lembrar: Capacidade de recordar informações. Questões que pedem definições ou fatos específicos;
- Entender: Capacidade de interpretar e explicar informações. Questões que pedem para resumir ou explicar conceitos;
- Aplicar: Capacidade de usar informações em contextos práticos. Questões que pedem para resolver problemas ou aplicar algoritmos;
- Analisar: Capacidade de decompor informações para entender suas partes. Questões que pedem para identificar padrões ou causas;
- Avaliar: Capacidade de julgar ou criticar informações com base em critérios. Questões que pedem para argumentar ou justificar uma posição;

- Criar: Capacidade de gerar algo novo, como ideias, produtos ou soluções. Questões que pedem para projetar, propor alternativas ou criar algoritmos.

Na pesquisa, essa hierarquia é utilizada para (i) orientar a geração de itens, (ii) verificar cobertura de habilidades cognitivas no conjunto de questões e (iii) auxiliar a avaliação (humana e automática). A taxonomia funciona como eixo de validação de conteúdo e alinhamento instrucional entre objetivos, itens e critérios de julgamento.

2.5 Critérios de avaliação das questões

Para assegurar a qualidade das questões geradas adotou-se um quadro multicritério que contempla propriedades formais, semânticas e pedagógicas de cada questão. Especificamente, serão avaliados: Cobertura (Nenkova; Passonneau, 2004), Gramaticalidade (Maity; Deroy; Sarkar, 2024a,b), Utilidade (Elkins *et al.*, 2023), Relevância (Mulla; Gharpure, 2023; Steuer *et al.*, 2021; Wang *et al.*, 2022) e Capacidade de resposta (Steuer *et al.*, 2021; Wang *et al.*, 2022) atribuindo pontuações em uma escala pré-determinada (1 a 5) para cada critério, como (Maity; Deroy; Sarkar, 2025) utilizou em seu experimento.

Detalhes dos critérios de avaliação:

- Cobertura: Este indicador é quantificado numa escala de 1 a 5, medindo a extensão em que as perguntas geradas abrangem os elementos essenciais do contexto fornecido. A cobertura é avaliada comparando as perguntas geradas às informações pertinentes dentro do contexto, especificamente identificando detalhes-chave que deveriam ser abordados. Uma pontuação mais alta reflete uma gama mais ampla de informações relevantes cobertas pelas perguntas, alinhando-se aos objetivos educacionais delineados no contexto.

Os critérios específicos para cada pontuação:

1: Poucos ou nenhum elemento relevante coberto do contexto.

2: Alguns elementos são cobertos de contexto, mas detalhes importantes estão ausentes.

3: Vários elementos importantes são cobertos do contexto, mas alguns detalhes são negligenciados.

4: Cobertura ampla dos elementos essenciais do contexto.

5: Cobertura abrangente de todos os elementos essenciais do contexto.

- Gramaticalidade: Este critério é avaliado numericamente numa escala de 1 a 5, focando na aderência das perguntas geradas às regras e convenções gramaticais padrão. Os avaliadores examinam a estrutura linguística, a sintaxe e a clareza das perguntas para determinar sua correção gramatical. Uma pontuação de 5 significa que a pergunta está livre de erros gramaticais, enquanto pontuações mais baixas refletem problemas variados de gramática. Os critérios específicos para cada pontuação:

1: Numerosos erros; difícil de entender.

2: Vários erros; o significado é parcialmente obscuro.

3: Alguns erros, geralmente compreensíveis.

4: Erros menores; maior parte clara e correta.

5: Sem erros; completamente clara e correta.

- Utilidade: A utilidade de uma pergunta é classificada numa escala de 1 a 5, indicando seu potencial valor para inclusão na avaliação de um professor. Essa métrica avalia se uma pergunta pode contribuir efetivamente para a aprendizagem dos estudantes e para os resultados de avaliação. É importante notar que uma pergunta não precisa ser totalmente respondível com base no contexto para ser considerada útil.

Os critérios específicos para cada pontuação:

- 1: Não útil; nenhuma contribuição para a aprendizagem.
 - 2: Utilidade limitada; impacto mínimo na avaliação.
 - 3: Moderadamente útil; contribui para a aprendizagem, embora seja um pouco previsível.
 - 4: Útil; melhora a avaliação e a compreensão do estudante.
 - 5: Altamente útil; aumenta significativamente a aprendizagem e os resultados da avaliação.
- Relevância: Este indicador avalia o alinhamento semântico entre cada pergunta gerada e o contexto de entrada correspondente, também avaliado numa escala de 1 a 5. Uma pergunta é considerada relevante se estiver diretamente relacionada ao conteúdo e aos temas apresentados no contexto. Os avaliadores consideram se a pergunta mantém o foco nos conceitos-chave e evita a introdução de ideias não relacionadas. Essa metodologia de avaliação é consistente com abordagens utilizadas em estudos anteriores, nos quais avaliações binárias e escaladas foram implementadas para medir.

Os critérios específicos para cada pontuação:

- 1: Não relevante; sem conexão com o contexto.
- 2: Pouco relevante; conexão fraca com o contexto.
- 3: Moderadamente relevante; algum foco no contexto.
- 4: Relevante; alinhada de perto com o contexto.
- 5: Altamente relevante; abordar diretamente o contexto sem ideias não relacionadas

- Capacidade de resposta: É avaliada numa escala de 1 a 5, usada para medir se uma pergunta pode ser respondida com base no contexto fornecido. Encontrar uma porção exata no contexto não é obrigatório; em invés disso, se um estudante puder responder razoavelmente à pergunta usando a lógica transmitida no contexto, ela é considerada "respondível".

Os critérios específicos para cada pontuação:

- 1: Não respondível; sem base no contexto.
- 2: Difícil de responder; conexão fraca com o contexto.
- 3: Moderadamente respondível; alguma lógica pode ser aplicada a partir do contexto.
- 4: Quase respondível; base clara no contexto.
- 5: Totalmente respondível; diretamente suportada pelo contexto.

2.6 Inteligência Artificial (IA)

Nesta dissertação a Inteligência Artificial (IA) é tratada no sentido de estudo de agentes inteligentes, sistemas que percebem o ambiente e escolhem ações para maximizar um objetivo (Russell; Norvig, 2004). O recorte técnico foca nas vertentes que dão suporte direto às próximas seções: Aprendizado de Máquina (ML), Aprendizado Profundo (DL) e Processamento de Linguagem Natural (PLN), conforme o esquema apresentado na Figura 2.

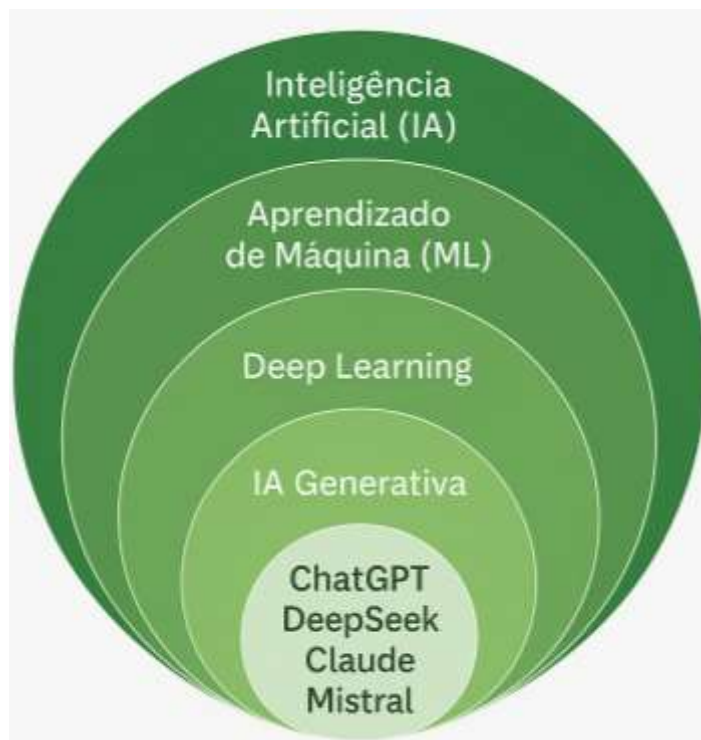


Figura 2 - Estrutura hierárquica da inteligência artificial, evidenciando a progressão de conceitos de ia geral até a especialização em modelos generativos.

Fonte: Elaborada pela autora (2025).

Como referência institucional, a AAI agrupa trabalhos em subáreas como Robótica, ML, Mineração de Dados/Big Data, Planejamento, Representação de Conhecimento, Raciocínio (probabilístico e simbólico), PLN, Sistemas (Multi)Agentes e Aplicações (Russell; Norvig, 2004, AAI).

No contexto educacional, a tecnologia viabiliza personalizar atividades, análise de dados de aprendizagem e geração automática de itens de avaliação. A literatura recente em Ética de IA na Educação reforça a necessidade de governança para privacidade, vieses e transparência, especialmente quando modelos são usados para produzir ou avaliar conteúdos instrucionais (Holmes *et al.*, 2022). Revisões nacionais e internacionais indicam adoção crescente, mas com ênfase em uso complementar ao trabalho docente (Gomes, 2010; Oliveira *et al.*, 2023).

2.7 Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural (PLN) é a subárea da IA dedicada a modelar computacionalmente a compreensão, a geração e a interação em

linguagem humana, oral e escrita, integrando linguística computacional, estatística e aprendizado de máquina (Jurafsky; Martin, 2023). Entre os métodos clássicos destacam-se a análise sintática (*parsing*), que modela a estrutura hierárquica e dependências gramaticais, e a análise semântica, que atribui significado considerando desambiguação lexical e contexto (Chowdhury, 2003; Jurafsky; Martin, 2023).

O avanço recente do PLN decorre de redes neurais profundas e, sobretudo, de Modelos de Linguagem de Larga Escala (LLMs), treinados em grandes corpora para capturar regularidades estatísticas e relações semântico-pragmáticas. Esses modelos viabilizam tarefas como tradução, sumarização, resposta a perguntas e geração condicionada de texto, frequentemente com desempenho de estado da arte (Jurafsky; Martin, 2023). No contexto educacional, o PLN tem sido aplicado à automação de atividades pedagógicas (p. ex., sumarização de materiais, apoio à autoria e elaboração de exercícios), com potencial para reduzir carga docente e ampliar personalização (Yang *et al.*, 2023).

Um caso específico é a Geração de Perguntas Educacionais (GPE), cujo objetivo é produzir automaticamente itens relevantes a partir de textos, imagens ou bases estruturadas (Zhang *et al.*, 2021). Diferentemente de tradução ou sumarização, a GPE exige identificar lacunas de informação (o que perguntar), tipos de pergunta e restrições linguísticas para garantir correção gramatical e adequação contextual (Junqueira; Corrêa; Freitas, 2024). Na educação, GPE é amplamente explorado para avaliação formativa e somativa e para composição de materiais didáticos (Kumar *et al.*, 2024). A combinação de técnicas de PLN com LLMs tem permitido *pipelines end-to-end* ou híbridos (seleção de respostas-alvo, geração controlada e verificação), incrementando cobertura e qualidade linguística (Junqueira; Corrêa; Freitas, 2024; Zhang *et al.*, 2021).

2.8 Modelos de Linguagem de Larga Escala (LLM)

LLMs são modelos de aprendizado profundo treinados em larga escala para modelagem de linguagem (previsão do próximo *token* ou reconstrução de *tokens* mascarados), capazes de generalizar para múltiplas tarefas com ou sem ajuste fino. A família *Transformer* sustenta esses modelos, permitindo atenção paralela e

captura de dependências de longo alcance (Amatriain *et al.*, 2024). Arquiteturas *encoder-only* (ex., *BERT*), *decoder-only* (ex., GPT) e *encoder-decoder* (ex., T5) diferem pelo objetivo de pré-treino e pelo modo de inferência, cobrindo desde compreensão contextual até geração controlada de texto. Na prática, modelos como GPT-3.5-turbo, *BERT* e T5 demonstram forte desempenho em geração, tradução, sumarização e QA, com ganhos substanciais após instrução e ajuste fino para domínios específicos (Amatriain *et al.*, 2024; Jurafsky; Martin, 2023).

Historicamente, o campo evoluiu de modelos *n-grama* para *Transformers* escalados, incorporando objetivos de pré-treino mais expressivos e dados massivos. Trabalhos como *BERT* ampliaram a compreensão contextual; *GPT* consolidou a geração autorregressiva em grande escala; e *T5* unificou tarefas de PLN sob o paradigma texto-para-texto, facilitando transferência entre tarefas (Amatriain *et al.*, 2024). Essa trajetória – frequentemente sintetizada em “árvores evolutivas” dos LLMs que pode ser vista na figura 3 – evidencia ciclos de escala, dados e otimização que expandem a capacidade para lidar com ambiguidade, raciocínio de múltiplos passos e adaptação a tarefas especializadas (Yang *et al.*, 2023).

relevância, estilo e formato da resposta. A eficácia do *Prompt* depende de clareza semântica, especificidade de instruções e estruturação do contexto; enunciados vagos tendem a induzir respostas imprecisas, ao passo que instruções explícitas e bem delimitadas favorecem comportamentos-alvo (ex., tom formal, concisão, formato de saída) (Brown *et al.*, 2020). Em LLMs, o *Prompt* explora aprendizado no contexto (*in-context learning*), pelo qual exemplos e instruções embutidos no próprio *Prompt* induzem o modelo a generalizar sem atualização de pesos (Brown *et al.*, 2020). Abaixo, serão apresentados os principais tipos de *Prompt*, suas definições e como eles impactam as respostas geradas.

Zero-Shot é uma abordagem na qual o usuário fornece apenas a tarefa ou pergunta ao LLMs, sem oferecer exemplos ou contexto adicional. O modelo é então responsável por entender a solicitação e gerar uma resposta baseada em seu conhecimento interno. É chamado de "*zero-shot*" porque o modelo não recebe nenhum exemplo prévio para embasar sua resposta. As respostas dependem diretamente do conhecimento do modelo sobre o tema abordado. Se o tema for muito específico ou complexo, a resposta pode ser genérica ou imprecisa. Por outro lado, se o tema for amplamente coberto no treinamento do modelo, a resposta tende a ser mais consistente (Kojima *et al.*, 2023).

Few-Shot é uma técnica que combina a tarefa ou pergunta com alguns exemplos de respostas desejadas. Esses exemplos ajudam o modelo a entender melhor o que é esperado, tornando as respostas mais direcionadas e consistentes. O termo "*few-shot*" se refere ao pequeno número de exemplos fornecidos, geralmente entre 1 a 5. Os exemplos fornecidos ajudam o modelo a compreender melhor o que é esperado, resultando em respostas mais personalizadas e coerentes. Além disso, os exemplos servem como guia para o tom, estilo e estrutura da resposta, aumentando a probabilidade de obter resultados satisfatórios (Brown *et al.*, 2020).

Chain-of-Thought é a abordagem na qual o usuário solicita que o modelo demonstre, passo a passo, o seu processo de raciocínio ao resolver uma tarefa ou responder a uma pergunta. Essa técnica é particularmente útil em contextos educacionais ou de depuração, onde entender o processo de tomada de decisão do modelo é tão importante quanto a resposta final (Wei *et al.*, 2023).

Role-Playing envolve instruir o modelo a assumir um papel ou personalidade específica. Essa abordagem é amplamente utilizada para personalizar as respostas, tornando-as mais coerentes com o contexto ou papel assumido. É uma técnica versátil que pode ser aplicada em uma variedade de cenários, desde interações educacionais até aplicações criativas (Kong *et al.*, 2024).

2.9.2 Meta – Prompt

Meta-Prompt é uma técnica de engenharia de *Prompt* de ordem superior na qual um primeiro *Prompt* especifica a forma, o esquema e as restrições de um *Prompt* de tarefa que será gerado em seguida; este *Prompt* derivado, guia a produção da resposta final pelo LLM, Tratando-se de um processo em duas etapas: $P_{meta} \Rightarrow P_{tarefa} \Rightarrow \text{saída}$. A ênfase recai na estrutura (*templates*, campos obrigatórios, tipos de entrada/saída) e na sintaxe (regras de formatação, marcadores, rubricas), desacoplando conteúdo e forma para ampliar controlabilidade, reuso e auditabilidade das interações (Zhang *et al.*, 2021).

Essa técnica é particularmente útil em contextos educacionais, onde professores podem usar prompts personalizados para criar exercícios e atividades para os alunos. Características principais:

- Orientado à estrutura: Modelar problemas e soluções com base em padrões, priorizando o formato.
- Focado na sintaxe: Usa a sintaxe como um guia para a geração de respostas.
- Exemplos abstratos: Emprega exemplos abstratos para descrever a morfologia de problemas e soluções.
- Versátil: Possui alta aplicabilidade entre domínios. É capaz de fornecer respostas estruturadas a uma ampla gama de problemas.
- Abordagem categórica: Baseia-se na teoria dos tipos para enfatizar a tipagem e a organização lógica dos *prompts*.

Na educação, o *Meta-Prompt* habilita geração sistemática de exercícios e

materiais a partir de esquemas didáticos (objetivo de aprendizagem, nível da Taxonomia de *Bloom*, habilidades BNCC, gabarito, rubrica, nível de dificuldade). Ao induzir o LLM a produzir primeiro o *Prompt* de tarefa com esses campos, o docente obtém itens mais padronizados, auditáveis e adaptáveis a turmas/perfis distintos. Evidências sobre personalização instrucional apontam ganhos de desempenho e engajamento quando o conteúdo é adaptado ao aluno/grupo (Pane *et al.*, 2015); o *Meta-Prompt* provê o mecanismo técnico para essa adaptação de forma escalável. Estudos também reportam ganhos na adaptação e personalização de conteúdo ao explicitar no *Prompt* as dimensões alvo (ex., critérios de avaliação, estilo de feedback) (Reynolds; McDonell, 2021).

3.1 Planejamento

A difusão de LLMs tem reconfigurado práticas educacionais, particularmente na geração de exercícios e conteúdos pedagógicos. Esta revisão analisa o estado da arte sobre a utilização de LLMs na educação, com foco específico na GPE e nas técnicas de *Meta-Prompt*. A metodologia escolhida para nortear este estudo foi a análise exploratória da literatura. Para esta revisão, foram realizadas pesquisas em cinco bases: *ACM*, *IEEE*, *Scopus*, *ScienceDirect* e *SpringerLink*, cobrindo o período a partir de 2019 marco temporal que coincide com avanços significativos em modelos como GPT-2 e subsequentes, que revolucionaram as capacidades de geração de texto contextualizado.

Para avaliar a elegibilidade dos estudos, foram estabelecidos critérios de inclusão e exclusão. Os critérios de inclusão determinam quais estudos podem ser considerados na pesquisa, mesmo após uma análise preliminar, enquanto os critérios de exclusão têm como objetivo filtrar e eliminar estudos que não atendem aos requisitos desejados. A Tabela 1 apresenta os critérios selecionados para esse processo.

Tabela 1 - Critérios de inclusão e exclusão utilizados.

	Critérios
Inclusão	Estudos que abordem a geração de perguntas educacionais
Exclusão	Estudos duplicados
Exclusão	Não respondem as perguntas de pesquisa
Exclusão	Não estar em inglês
Exclusão	Estudos não disponíveis na íntegra
Exclusão	Estudos anteriores a 2019

Fonte: Elaborada pela autora (2025).

3.2 Investigação do problema

Visando entender como o assunto está sendo abordado na academia, foram elaboradas perguntas de pesquisa secundárias para se extrair as informações necessárias dentro dos estudos encontrados.

Q1 Quais métodos de avaliação são mais frequentemente utilizados para validar a qualidade de questões educacionais geradas por LLMs na literatura?

Q2 Como a evolução temporal dos modelos de LLM (2019-2024) impactou as abordagens e resultados reportados na geração de questões educacionais?

Q3 Quais teorias de aprendizagem fundamentam as abordagens de GPE com LLMs na literatura revisada?

Q4 Quais desafios de implementação prática são mais frequentemente mencionados para adoção de GPE com LLMs em contextos reais?

1. Métodos de Avaliação na GPE com LLMs

A literatura examinada demonstra consenso quanto à necessidade de abordagens multimétodo para avaliação da qualidade das questões geradas. Predomina o uso de validação híbrida, combinando métricas técnicas com avaliação humana especializada. Estudos como (Elkins *et al.*, 2023; Maity; Deroy; Sarkar, 2025) empregaram sistematicamente a avaliação por professores, utilizando escalas *Likert* e critérios baseados na Taxonomia de *Bloom*. Maity (2025) especificamente recrutou professores com dez ou mais anos de experiência através da plataforma *UpWork*, estabelecendo critérios rigorosos de seleção baseados em experiência docente e familiaridade com a disciplina.

Paralelamente, métricas técnicas de avaliação incluem relevância semântica [Wang, X. *et al.* 2022], correção gramatical (Elkins *et al.*, 2023) e cobertura de conteúdo (Maity; Deroy; Sarkar, 2025). A evolução metodológica mostra uma transição de *benchmarks* automatizados, como o uso do *dataset SQuAD* (Chan; An; Davoudi, 2023), para protocolos de validação mais sofisticados que incorporam dimensões pedagógicas e contextuais. Esta trajetória reflete o amadurecimento do campo, que progressivamente transcende a avaliação puramente técnica para abraçar critérios educacionalmente significativos.

2. Evolução Temporal dos Modelos e Impacto na GPE

A análise dos estudos revela uma correlação positiva entre a evolução tecnológica dos LLMs e a sofisticação das abordagens de GPE. O período 2019-2021 (He *et al.*, 2022), marcado pela emergência de modelos como *GPT-2* e aplicações iniciais do *BERT*, caracterizou-se por investigações exploratórias

focadas na viabilidade técnica básica. Nesta fase, as taxas de aceitação por especialistas situavam-se entre 45% e 60%, refletindo limitações significativas na qualidade pedagógica das questões geradas.

O período 2022-2023 avançou substancialmente com a adoção de GPT-3, GPT-3.5 e *text-davinci-003* (Tran *et al.*, 2023). Estudos como (Wang *et al.*, 2022)] demonstraram que estratégias de *Prompt* mais refinadas, incluindo contextos de entrada mais curtos e aprendizado com poucos exemplos, resultavam em perguntas de qualidade superior. As taxas de aceitação elevaram-se para 65%-75%, indicando melhorias mensuráveis na adequação pedagógica.

A partir de 2024, com a consolidação do *GPT-4* [Tran *et al.* 2023] e modelos subsequentes, os estudos se direcionaram para desafios mais complexos, incluindo o controle fino da geração por meio de *Meta-Prompt* (Elkins *et al.*, 2024) e a abordagem sistemática de vieses (Gallegos *et al.*, 2024; Weissburg *et al.*, 2025). Esta fase atual registra taxas de aceitação que superam 80%, acompanhadas por uma preocupação crescente com dimensões éticas e de equidade.

3. Fundamentos Teóricos das Abordagens de GPE

A análise revela que as abordagens de GPE com LLMs fundamentam-se em uma estrutura teórica diversificada. O Conectivismo (Siemens, 2004) fornece a base para compreender os LLMs como ferramentas de mediação em redes de aprendizagem digital, enquanto o Construcionismo (Papert, 1993) informa estratégias que enfatizam a construção ativa do conhecimento através da geração de questões. A Teoria da Carga Cognitiva (Sweller, 1994) emerge como referencial importante para o design de questões que otimizam o processamento de informação, evitando a sobrecarga cognitiva nos estudantes. Adicionalmente, elementos do Cognitivismo manifestam-se na adoção generalizada de taxonomias de aprendizagem, particularmente a Taxonomia de *Bloom* (Elkins *et al.*, 2023), para estruturar objetivos educacionais e níveis de complexidade cognitiva. Nota-se, contudo, que a fundamentação teórica frequentemente permanece secundária em relação às considerações técnicas, sugerindo uma lacuna significativa na articulação explícita entre pressupostos teóricos e implementações práticas.

4. Desafios de Implementação Prática

Os estudos revisados identificam desafios inter-relacionados que impactam a implementação da GPE com LLMs em contextos educacionais reais. Os desafios técnicos incluem a dependência crítica de *Prompts* bem elaborados (Wang *et al.*, 2022), que representa uma barreira significativa para educadores não familiarizados com engenharia de *Prompts*, e a persistência de vieses sistemáticos (Gallegos *et al.*, 2024; Weissburg *et al.*, 2025) que comprometem a equidade educacional. No domínio pedagógico, destaca-se a necessidade de validação humana contínua (Tran *et al.*, 2023), que impõe custos significativos de tempo e recursos especializados, e os desafios de adaptação cultural e contextual (Sasani, 2024; Sociedade, 2025). Estes últimos manifestam-se particularmente na tendência dos LLMs em privilegiar referências culturais ocidentais, limitando a relevância em contextos educacionais diversificados.

Os desafios institucionais incluem a resistência dos docentes decorrente de lacunas na formação tecnológica e preocupações com a autonomia profissional, além de questões de infraestrutura e governança (Lee *et al.*, 2024) relacionadas à implementação em escala.

A análise integrada demonstra que o campo da GPE com LLMs evoluiu significativamente em termos técnicos, com melhorias mensuráveis na qualidade das questões geradas. No entanto, persistem desafios substanciais relacionados à validação pedagógica, equidade e implementação prática.

3.3 Estudos de Referência

Por exemplo, (Wang *et al.*, 2022) realizaram experimentos com estratégias de *Prompt* para otimizar a GPE, concluindo que contextos de entrada mais curtos e aprendizado com poucos exemplos resultam em perguntas de maior qualidade. Técnicas de geração automática de perguntas (QG) impulsionadas por PNL têm potencial pedagógico, economizando tempo dos educadores e beneficiando a aprendizagem dos alunos. No entanto, essas técnicas ainda não são amplamente adotadas em salas de aula. Os autores defendem que o desenvolvimento de sistemas colaborativos de QG, que priorizem o controle do instrutor e a capacidade de explicação, é essencial para a adoção real dessas ferramentas. Elkins (2023)

investigaram a GPE em taxonomias pedagógicas, como a Taxonomia de *Bloom*, e, embora demonstrem o potencial da geração de texto controlável por meio de LLMs para transformar a educação, permaneceu incerto se os professores consideram essas perguntas úteis e de qualidade para o ambiente de sala de aula. Para essa lacuna, os autores realizaram uma avaliação humana com professores, combinando geração de texto controlável com a Taxonomia de *Bloom* e a dificuldade das questões. Os resultados indicaram que as perguntas geradas foram julgadas de alta qualidade e suficientemente úteis, sugerindo seu potencial para uso generalizado em contextos educacionais.

Chan (2023) investigou as capacidades do *ChatGPT* na tarefa de GPE, verificando seu desempenho em conjuntos de dados como o *SQuAD* e o manual de automóveis. Os resultados indicam que, com *Prompts* adequadamente projetados, o *ChatGPT* pode competir ou até superar modelos baseados em QG. Além disso, o estudo explorou o potencial de aprimoramento do modelo por meio de ajustes finos nos *Prompts*. Por fim, foi avaliado o uso do *ChatGPT* como ferramenta para avaliar a precisão de modelos de QG, demonstrando que seu desempenho pode ser comparável ao de avaliadores humanos, apesar de ainda estar em estágios iniciais.

Tran (2023) investigam o uso do GPT-3 e do GPT-4, para a geração automática de questões de múltipla escolha (MCQs) em cursos de computação. A criação de MCQs é uma tarefa desafiadora e demorada para professores, frequentemente recorrendo-se a bancos de questões existentes ou reutilizando materiais de semestres anteriores. Os autores empregaram técnicas de *Prompt engineering*, desenvolvendo três tipos de *Prompts* para garantir respostas corretas e distratores plausíveis, além de utilizarem o modelo "*text-davinci-003*" com temperatura padrão para controlar a variedade das respostas. Os resultados indicam que, especialmente o GPT-4, é uma ferramenta promissora para a geração de MCQs, permitindo a criação rápida de questões personalizadas. No entanto, ressalta-se que os modelos ainda não substituem a revisão humana, pois podem gerar respostas corretas como distratores ou produzir alternativas ambíguas.

A eficácia das avaliações educacionais é diretamente influenciada pela qualidade e pela diversidade das perguntas utilizadas, que desempenham um

papel fundamental na avaliação da compreensão dos alunos (Howard; Aleman, 2008). É essencial que os educadores garantam que as perguntas estejam adaptadas ao nível de conhecimento dos estudantes, ao mesmo tempo em que abordem questões-chave relevantes. Além disso, é importante que as perguntas estimulem o desenvolvimento de habilidades cognitivas de maior complexidade, como a análise, a aplicação e a avaliação, em vez de se concentrar apenas na memorização de fatos (Billing, 2007).

Os estudos sugerem que os LLMs têm o potencial de reduzir a carga de trabalho dos professores e personalizar o aprendizado, mas destacam a importância da validação humana e da qualidade dos *prompts*. Essas limitações reforçam a necessidade de pesquisas que avaliem a experiência de professores no uso de LLMs para elaboração de exercícios, complementando a supervisão humana com técnicas adequadas (Tran *et al.*, 2023; Wang *et al.*, 2022).

Os autores (Elkins *et al.*, 2024) definem especialistas como professores com experiência em ensino ou tutoriais de Biologia. Esses pré-requisitos foram considerados na seleção dos especialistas. No entanto, o artigo não menciona explicitamente a experiência comprovada dos especialistas na Taxonomia de Bloom. Por outro lado, os autores (Maity; Deroy; Sarkar, 2025) especificam os critérios de seleção dos professores envolvidos na avaliação das perguntas geradas pelos modelos de linguagem. Eles recrutaram professores de ensino médio através da plataforma *UpWork*, selecionando candidatos com base em sua experiência em educação de ensino médio, familiaridade com a matéria e um mínimo de dez anos de experiência docente. Esses educadores foram responsáveis por avaliar as perguntas feitas tanto por humanos quanto pelos modelos de linguagem, utilizando a escala *Likert* de avaliação de 1 a 5, com base em métricas pré-definidas, incluindo cobertura, gramaticalidade, praticidade, relevância e capacidade de resposta.

A GQE por LLMs traz oportunidades de escala e personalização, mas gera riscos para a igualdade em contextos multiculturais. Pesquisas mostram que vieses aparecem de forma sistemática na produção de conteúdos, respostas e avaliações, independentemente do uso específico (Gallegos *et al.*, 2024; Lee *et al.*, 2024; Sociedade, 2025; Weissburg *et al.*, 2025). Revisões sobre vieses e justiça em LLMs

sugerem que os modelos tendem a reproduzir ou ampliar desigualdades entre grupos demográficos, o que exige cuidado na aplicação educativa (Gallegos *et al.*, 2024). Em termos práticos, estudos mostram que os itens podem ter problemas na formulação, no uso de contextos culturais e na atribuição de significado a identidades, o que afeta a representatividade e a eficácia do ensino (Kotek; Dockum; Sun, 2023; Sasani, 2024; Sociedade, 2025; Weissburg *et al.*, 2025).

Os LLMs frequentemente exibem viés cultural em favor de referências ocidentais, mesmo quando solicitados a responder em línguas ou contextos não-ocidentais, o que pode levar à priorização de conteúdos (Weissburg *et al.*, 2025). Estudos indicam que a dominância de estruturas culturais ocidentais pode afetar a pertinência de itens gerados para estudantes de outras realidades culturais, reduzindo a identificação com o material e o engajamento (Sociedade, 2025; Weissburg *et al.*, 2025). Pesquisas adicionais destacam que a linguagem dominante (em particular o inglês) pode condicionar a seleção de conteúdos e exemplos, ignorando perspectivas culturais não americanas (Sasani, 2024). Essa tendência de centralidade cultural se reflete na formulação de questões que privilegiam contextos estrangeiros ou irrelevantes para determinado público estudantil, contribuindo para exclusão pedagógica e menor efetividade do ensino (Sasani, 2024; Sociedade, 2025).

Vieses socioeconômicos emergem quando conteúdos gerados por LLMs reproduzem estereótipos ou narrativas que podem desvalorizar grupos de baixa renda, ou apresentar tais populações sob ângulos estereotipados. Pesquisas apontam a ocorrência de representações que associam determinadas condições socioeconômicas a descrições carregadas de preconceitos, o que pode se traduzir em itens que reforçam estigmatização em avaliações e materiais pedagógicos (Silva Costa *et al.*, 2025; Sociedade, 2025; Weissburg *et al.*, 2025). Quando inseridos na geração de questões, esses vieses podem desmotivar estudantes de contextos socioeconômicos menos favorecidos, diminuindo o engajamento, a autoestima e a participação, com impactos diretos na equidade de aprendizado (Silva Costa *et al.*, 2025; Sociedade, 2025). Além disso, revisões sobre o ciclo de vida dos LLMs em educação destacam que atualizações e práticas de moderação podem, inadvertidamente, perpetuar ou ampliar tais representações distorcidas se

não houver governança e validação contínuas (Lee *et al.*, 2024).

Esta pesquisa busca preencher lacunas importantes identificadas na literatura sobre o uso de LLMs e *Meta-Prompt* na educação. Embora haja um crescente número de estudos sobre a GPE por LLMs, a perspectiva dos docentes sobre a utilidade e a qualidade dessas ferramentas ainda é pouco explorada. A maioria dos trabalhos foca na capacidade técnica dos LLMs, sem aprofundar na experiência e nas necessidades dos professores no contexto real da sala de aula.

4 METODOLOGIA

Este capítulo detalha o percurso metodológico empregado para investigar a qualidade de questões educacionais geradas por LLMs e por uma professora. O foco da análise recai sobre critérios como cobertura, gramaticalidade, relevância, utilidade e capacidade de resposta. O objetivo principal é explicitar as etapas de seleção dos LLMs, bem como a estruturação da coleta, geração e avaliação dos dados. A Figura 4 oferece uma visão sintética do fluxo seguido neste estudo, servindo como referência visual para as etapas descritas a seguir.

Estrutura proposta:

- 4.1 Seleção e teste dos LLMs apropriados: Esta etapa envolve a delimitação do conjunto de modelos a serem comparados e a realização de uma verificação preliminar de seu funcionamento e aderência aos requisitos do estudo, como idioma, limites de contexto e consistência de saída. O objetivo é estabelecer um *baseline* técnico que garanta a comparabilidade entre os modelos em condições controladas.
- 4.2 Refinamento dos *Prompts* para geração das questões: Nesta fase, define-se e padroniza-se o conjunto de *Prompts* a ser utilizado na geração das questões. Busca-se um equilíbrio entre clareza, controle de variáveis e alinhamento pedagógico, visando reduzir ambiguidades e vieses de formulação, o que contribui para a qualidade e consistência das saídas analisadas.
- 4.3 Avaliação das questões geradas: Esta etapa organiza o procedimento de avaliação, integrando instrumentos e critérios previamente definidos. As respostas são analisadas segundo métricas comparáveis entre os modelos e um *baseline* humano, permitindo a sumarização estatística e interpretações alinhadas às perguntas de pesquisa.

Processo de Avaliação de LLM



Figura 4 – Processo de avaliação de LLMs.

Nota: A sequência de etapas em azul corresponde à subseção 4.1, em rosa à subseção 4.2 e em laranja à subseção 4.3.

Fonte: Elaborada pela autora (2025).

4.1 Seleção e teste dos LLMs apropriados

A escolha dos LLMs para este estudo representa uma etapa crucial na investigação proposta. A seleção criteriosa dos LLMs é fundamental para investigar se a técnica de *Meta-Prompt* influencia a qualidade das questões geradas, impactando assim a eficácia do processo de avaliação educacional. Essa etapa é crucial pois a escolha de LLMs com diferentes arquiteturas e capacidades permite avaliar a robustez do *Meta-Prompt* em diversos cenários, garantindo que os resultados não sejam enviesados por características específicas de um único modelo.

Foi realizada uma análise dos LLMs disponíveis no mercado, cujos detalhes são apresentados na Tabela 2. Essa tabela oferece uma comparação dos principais

LLMs, destacando suas áreas de especialização, acessibilidade (incluindo a presença de um site oficial) e funcionalidades, com o intuito de facilitar a escolha da ferramenta mais adequada para a geração de questões educacionais. As informações foram coletadas a partir de pesquisas recentes e de documentação oficial, buscando refletir o estado atual da tecnologia na área de LLMs. Para a seleção dos modelos utilizados neste estudo, foram considerados os seguintes critérios, cada um com um papel fundamental para a validade e aplicabilidade da pesquisa:

- Disponibilidade de um site oficial, evitando a necessidade de implementação própria (garantindo a acessibilidade e facilidade de uso);
- Gratuidade do acesso (tornando a pesquisa replicável e acessível a outros pesquisadores);
- Capacidade de análise de documentos fornecidos pelo usuário (permitindo a avaliação do desempenho dos LLMs em diferentes tipos de materiais).

Tabela 2 - Comparação entre os LLMS.

LLM	Especialidade	Site	Gratuito	Análise de Documentos
<i>ChatGPT-4o</i>	Processamento multimodal, textos longos, criação de conteúdo extenso.	Sim	Parcial	Sim
<i>Claude 3.5</i>	Geração de texto criativa, raciocínio crítico, fluência quase humana.	Sim	Parcial	Sim
<i>DeepSeek-R1</i>	Raciocínio lógico, tomada de	Sim	Parcial	Sim

	decisões em tempo real.			
<i>Mistral AI</i>	Modelos eficientes e adaptáveis, bom equilíbrio entre qualidade e custo-benefício.	Sim	Código Aberto / Parcial	Sim
<i>Gemini</i>	Suporte a escrita, programação e criatividade com aprendizado de máquina.	Sim	Parcial	Sim
<i>Ernie</i>	Compreensão de texto em diferentes idiomas e contextos culturais. Não é traduzido	Sim	Sim	Não Confirmado
<i>Falcon</i>	Análise de segurança cibernética, monitoramento de ameaças em tempo real.	Sim	Não	Não Confirmado
<i>Guanaco</i>	Tarefas de linguagem com contextualização e interpretação.	Não	Sim	Sim
<i>Bloom</i>	Práticas inclusivas e diversidade em aplicações de inteligência artificial.	Não	Sim	Sim
<i>Alpaca</i>	Análise de comportamentos em ecossistemas agrícolas e biológicos.	Não	Sim	Sim
<i>LLaMA</i>	Geração de texto de alta qualidade, respostas eficazes, ajustável.	Não	Código Aberto	Sim

<i>Gemma</i>	Serviços de IA em linguagem natural e soluções generativas personalizadas.	Não	Código Aberto	Não Confirmado
<i>OpenLLaMA</i>	Geração de linguagem de código aberto, ajustável para várias aplicações.	Não	Código Aberto	Sim
<i>Dolly 2.0</i>	Criação de modelos personalizados com base em texto e instruções.	Não	Código Aberto	Sim

Fonte: Elaborada pela autora (2025).

Após a análise comparativa das opções disponíveis no mercado, e considerando os critérios estabelecidos de acessibilidade, gratuidade e capacidade de análise de documentos, a seleção final recaiu sobre cinco ferramentas: *Gemini*, *DeepSeek*, *Claude*, *Mistral* e *ChatGPT*.

O objetivo do teste preliminar é avaliar a eficácia dos LLMs selecionados, com ênfase na capacidade de gerar GPE de qualidade, especialmente quando empregadas técnicas de *Meta-Prompt*. A avaliação concentrou-se nos seguintes aspectos:

- Clareza: Determinar se os exercícios gerados pelos LLMs, utilizando *Meta-Prompt*, são claros, precisos e facilmente compreensíveis para os alunos.
- Nível de dificuldade apropriado: Avaliar se os exercícios possuem um nível de dificuldade adequado ao público-alvo e aos objetivos de aprendizagem.
- Alinhamento com o currículo: Avalie em que medida os exercícios gerados estão alinhados com o currículo e com os objetivos de aprendizagem específicos.

- Ausência de erros: Certifique-se de que os exercícios estejam livres de erros, e forneçam respostas corretas e inequívocas.

Conduziu-se uma checagem exploratória com os cinco LLMs selecionados, aplicando o mesmo *Prompt* e o mesmo *Meta-Prompt* descritos no Quadro 1. As saídas foram registradas e organizadas nas Tabelas A.1 e A.3 (Apêndice A). O objetivo desta etapa foi verificar a conformidade de formato, níveis de dificuldade, clareza do enunciado e ausência de erros evidentes.

Quadro 1 - *Prompt* e *Meta-Prompt* usados nos testes preliminares.

<i>Prompt</i> : Crie cinco questões sobre Língua Portuguesa para alunos do sétimo ano. Essas questões devem seguir os níveis da Taxonomia de <i>Bloom</i> .
<i>Meta-Prompt</i> : Crie um <i>Prompt</i> para criar cinco questões sobre Língua Portuguesa para alunos do sétimo ano. Essas questões devem seguir os níveis da Taxonomia de <i>Bloom</i> .

Fonte: Elaborada pela autora (2025).

Esta seção apresenta uma análise comparativa de cinco LLMs quanto à capacidade de gerar questões educacionais alinhadas à Taxonomia de *Bloom*. Para tal, aplicaram-se um *Prompt* e um *Meta-Prompt* a fim de avaliar a precisão, consistência e flexibilidade das saídas. O objetivo é identificar particularidades, limitações e o potencial de cada modelo na elaboração de itens educacionais; as especificidades observadas são descritas a seguir.

Na análise da Tabela A.1, que se encontra no apêndice A, todos os cinco LLMs (*Gemini*, *DeepSeek*, *Claude*, *Mistral* e *ChatGPT*) demonstraram precisão ao gerar cinco questões alinhadas com a Taxonomia de *Bloom*, formuladas de maneira clara, direta e adequada para estudantes do 7º ano. *Claude* e *ChatGPT* destacaram-se pela inclusão de uma variedade de formatos, como múltipla escolha e dissertativa, enriquecendo o potencial avaliativo. A quinta questão de *Claude*, em particular, combinou dois níveis da Taxonomia de *Bloom*, evidenciando uma

interpretação criativa e válida do referencial teórico.

Em termos de consistência, todos os LLMs rotularam de forma clara os níveis da Taxonomia de *Bloom* em suas questões. Contudo, *Claude* e *ChatGPT* apresentaram variações no formato das questões geradas, o que, embora adicione variedade, pode ser um ponto a ser considerado dependendo dos requisitos de padronização. Quanto à flexibilidade, os cinco LLMs abordaram uma ampla gama de tópicos em Língua Portuguesa, adaptando a complexidade das perguntas aos diferentes níveis da Taxonomia de *Bloom* para assegurar uma avaliação abrangente. Uma observação relevante é que *Claude* não incluiu uma imagem ou referência visual em sua quinta questão, o que poderia ter potencializado o engajamento ou a compreensão.

Considerando os demais LLMs, *Gemini*, *DeepSeek* e *Mistral* cumpriram as exigências de clareza e pertinência na elaboração das questões, mas não apresentaram a mesma variedade de formatos observada em *Claude* e *ChatGPT*. Dentre esses, o *Mistral*, ao manter um formato mais tradicional e homogêneo, posicionou-se ligeiramente abaixo dos modelos que exploraram uma abordagem mista. Assim, embora todos tenham demonstrado desempenho satisfatório, o *ChatGPT* emergiu como o modelo com a avaliação mais robusta, enquanto o *Mistral* ficou um pouco aquém em termos de inovação na apresentação das questões.

Na análise da Tabela A.3, que se encontra no apêndice A, os cinco LLMs geraram questões que cobriram os níveis da Taxonomia de *Bloom*. *Gemini* produziu cinco questões sobre o tópico "substantivos", conforme especificado pelo *Meta-Prompt*. Contudo, a questão 3 apresentou inconsistência na formatação das opções, o que pode gerar confusão. *DeepSeek*, por sua vez, excedeu o escopo ao incluir uma questão adicional e falhou em explicitar a resposta correta na questão 1 (objetiva).

Claude demonstrou alta precisão ao seguir a estrutura detalhada do *Meta-Prompt*, utilizando formatos de múltipla escolha e dissertativa, e incluindo a resposta esperada com uma justificativa pedagógica para cada questão. De forma similar, *ChatGPT* aderiu fielmente à estrutura do *Meta-Prompt*, apresentando

habilidade cognitiva, enunciado, opções para questões objetivas e critérios de correção para as dissertativas. Em contraste, o *Mistral* não seguiu o formato detalhado de múltipla escolha com gabarito e justificativa, optando por manter questões dissertativas, o que desviou do objetivo do *Meta-Prompt*.

Em termos de consistência, *Gemini* e *DeepSeek* mantiveram um formato majoritariamente consistente e rotularam os níveis da Taxonomia de *Bloom*, fornecendo a opção correta. No entanto, o *DeepSeek* apresentou variação no formato (objetiva/dissertativa), comprometendo a uniformidade. *Claude* e *ChatGPT* exibiram um formato de apresentação consistente para cada questão, incorporando todos os elementos solicitados pelo *Meta-Prompt*, como habilidade cognitiva, enunciado, opções e gabarito/justificativa. O *Mistral*, embora consistente em seu formato dissertativo, mostrou-se inconsistente com o formato mais estruturado e objetivo implícito no *Meta-Prompt*, conforme interpretado pelos outros LLMs.

Quanto à flexibilidade, o *Gemini* demonstrou capacidade de criar diferentes tipos de perguntas (definição, identificação, aplicação, análise, avaliação) porém somente dentro do tópico "substantivos" e do formato de múltipla escolha. O *DeepSeek* exibiu boa flexibilidade ao abordar diversos conteúdos (conjunções, figuras de linguagem, concordância verbal, variação linguística, produção textual) e ao combinar formatos de questão, incluindo um nível bônus ("Criar"). O *Claude* também mostrou flexibilidade ao elaborar questões complexas e detalhadas para cada nível, cobrindo uma variedade de conteúdos (classes gramaticais, figuras de linguagem, coesão textual) e fornecendo justificativas pedagógicas. O *ChatGPT* destacou-se pela excelente flexibilidade, criando questões em múltiplos formatos (múltipla escolha e dissertativa) e abordando diversos conteúdos (ortografia, interpretação, pontuação, figuras de linguagem, criação de texto), com detalhes sobre a correção. O *Mistral*, embora flexível na cobertura de conteúdos (classes de palavras, figuras de linguagem, análise sintática, criação de diálogo), não adaptou seu formato de saída para seguir as instruções mais específicas do *Meta-Prompt*, limitando sua flexibilidade no que tange ao formato de múltipla escolha com gabarito e justificativa.

Em síntese, o *ChatGPT* obteve a melhor avaliação, seguindo com exatidão as orientações do *Meta-Prompt* e produzindo questões em formatos variados, com

respostas esperadas e justificativas pedagógicas detalhadas. Esse rigor garantiu uma abordagem completa e consistente. Em contrapartida, o *Mistral* apresentou a avaliação mais baixa, pois, apesar de abordar diversos conteúdos, optou por um formato exclusivamente dissertativo, divergindo das instruções mais elaboradas do *Meta-Prompt*. *Gemini* demonstrou pouca flexibilidade ao focar apenas em "substantivos", enquanto o *DeepSeek*, apesar da variedade de conteúdos e formatos, excedeu o escopo e omitiu a resposta correta em uma questão. O *Claude* manteve uma estrutura consistente e sólida, mas sem o nível de aprofundamento do *ChatGPT*.

Durante a fase dos testes preliminares, foi repetido em duas ocasiões adicionais. Embora o conteúdo das questões tenha sido alterado e os formatos variados, as mesmas tendências e variações observadas inicialmente foram mantidas. Em uma das repetições, o *Claude* gerou uma questão com referência a uma imagem, mas sem o fornecimento da mesma, o que comprometeu a compreensão. Este detalhe, embora pontual, foi um fator decisivo para a exclusão do *Claude* no experimento principal. Esses testes foram fundamentais para nortear o desenvolvimento do experimento principal. Com base nos resultados obtidos, foi possível realizar alterações precisas tanto no texto do *Prompt* quanto no *Meta-Prompt*, estabelecendo de forma mais detalhada as expectativas quanto à variação e à estrutura das questões. Adicionalmente, a BNCC foi incorporada como referência central, garantindo que o conteúdo das questões estivesse alinhado com as diretrizes curriculares nacionais e padronizado para atender aos objetivos pedagógicos. Diante dessa análise e dos ajustes realizados, um novo teste foi conduzido com um *Meta-Prompt* revisado que pode ser visto no quadro 2, cujas respostas geradas por cada ferramenta estão apresentadas nas Tabelas A.4 no Apêndice B.

Quadro 2 - Meta-Prompt revisado para o teste.

Meta-Prompt: Você é um especialista em educação com profundo conhecimento da Base Nacional Comum Curricular (BNCC) de Língua Portuguesa e da Taxonomia de Bloom. Sua tarefa é criar um *prompt* especializado que será usado posteriormente para gerar 3 questões educacionais de alta qualidade para alunos do 7º ano do Ensino Fundamental. As questões devem incluir contextos significativos e relevantes para estudantes desta faixa etária, enunciados claros e objetivos e podem ser em qualquer formato.

Fonte: Elaborada pela autora (2025).

No caso do *ChatGPT*, o *Meta-Prompt* original não especificava a necessidade de gerar múltiplas questões, resultando inicialmente na produção de apenas uma. Para atingir o objetivo de gerar três questões, foi essencial ajustar o *Prompt*, explicitando essa demanda. Adicionalmente, a formatação em *Markdown* precisou ser removida para adequar o resultado ao formato desejado. Similarmente, o *Mistral* gerou apenas uma questão na primeira tentativa, devido à falta de clareza no *Meta-Prompt* quanto à quantidade de questões. A questão produzida replicou a estrutura do exemplo fornecido, o que limitou a diversidade.

Esses testes foram cruciais para o desenvolvimento do experimento, pois permitiram identificar as melhores práticas na construção de questões e evidenciaram a importância de ajustar os instrumentos de avaliação. Com base nos resultados, foram realizadas alterações precisas tanto no *Prompt* quanto no *Meta-Prompt*, detalhando as expectativas de variação e estrutura das questões. A incorporação da BNCC como referência central garantiu o alinhamento do conteúdo das questões com as diretrizes curriculares nacionais e a padronização para os objetivos pedagógicos.

4.2 Refinamento dos *Prompts* para geração das questões

Os *prompts* foram elaborados com base nas avaliações realizadas posteriormente aos testes preliminares. O experimento foi conduzido utilizando as quatro LLMs (*ChatGPT*, *DeepSeek*, *Mistral* e *Gemini*), aplicando o mesmo *Prompt*

e *Meta-Prompt* que podem ser vistos no quadro 3 para a criação das questões. As respostas geradas por cada LLMs estão apresentadas nas tabelas B1 a B4 no apêndice B:

Quadro 3 - Prompt, Meta-Prompt e pedido para a professora para criarem as questões para o experimento.

Professor: Crie três questões com o conteúdo definido na BNCC sobre Língua Portuguesa para alunos do sétimo ano. Essas questões devem seguir os níveis da Taxonomia de <i>Bloom</i> . As questões podem ser em qualquer formato.
<i>Prompt</i> : Crie três questões sobre Língua Portuguesa para alunos do sétimo ano do Ensino Fundamental, seguindo o conteúdo definido na Base Nacional Comum Curricular (BNCC) e os diferentes níveis da Taxonomia de Bloom. As questões devem incluir contextos significativos e relevantes para estudantes desta faixa etária, enunciados claros e objetivos e podem ser em qualquer formato.
<i>Meta-Prompt</i> : Você é um especialista em educação com profundo conhecimento no que diz respeito à Base Nacional Comum Curricular (BNCC) de Língua Portuguesa e sempre aplica os conceitos da Taxonomia de Bloom na criação de seu material. Sua tarefa é criar um <i>Prompt</i> que será usado para gerar 3 questões educacionais de alta qualidade para alunos do 7º ano do Ensino Fundamental. As questões devem incluir contextos significativos e relevantes para estudantes desta faixa etária, enunciados claros e objetivos e podem ser em qualquer formato. As questões já devem vir formatadas para serem aplicadas ao documento da prova.

Fonte: Elaborada pela autora (2025).

Durante a análise inicial dos dados coletados na GPE, identificou-se uma característica importante nos resultados da LLMs *Gemini*. Observou-se que as questões criadas pelo *Gemini* usando o *Prompt* tinham uma grande semelhança, sendo em alguns casos quase idênticas aos exemplos de questões que foram fornecidos na descrição do *Meta-Prompt*. Além disso, testes preliminares

mostraram que o *Gemini* demonstrou pouca flexibilidade, elaborando perguntas focadas apenas no tema "substantivos", mesmo quando seu *Meta-prompt* aborda outros assuntos. Essa limitação levantou dúvidas sobre sua capacidade de seguir instruções variadas de forma consistente no experimento

A semelhança das questões do *Gemini* com os exemplos do *Meta-Prompt* gerou uma preocupação significativa quanto à validade da comparação entre os dois tipos de prompt para este LLM. A explicação mais provável é que o *Gemini*, ao usar o *Prompt*, não tenha gerado as questões de forma independente, mas sim copiado ou se baseado exclusivamente nos exemplos do *Meta-prompt*, possivelmente devido a informações acessadas durante seu treinamento. Tal comportamento comprometeria a comparação direta entre o desempenho do *Prompt* e do *Meta-prompt* para o *Gemini*, introduzindo uma distorção nos resultados do experimento. Por esses motivos, e para manter a validade do estudo e a confiabilidade das comparações entre as estratégias de *Prompt*, decidiu-se excluir os dados do *Gemini* do questionário.

4.3 Avaliação das questões geradas

A qualidade das questões será avaliada por meio de duas abordagens: avaliação automática e avaliação realizada por especialistas.

Na avaliação humana, será solicitado por e-mail que um professor experiente na disciplina de Língua Portuguesa analise as questões criadas e responda se elas estão de acordo com os critérios pré-definidos (Cobertura, Gramaticalidade, Utilidade, Relevância e Capacidade de resposta) atribuindo pontuações em uma escala pré-determinada (1 a 5) para cada critério, como (Maity; Deroy; Sarkar, 2025) utilizou em seu experimento. Se as questões não forem aceitas, o processo de geração das questões será repetido; se foram aceitas, um questionário será montado (apêndice C) e o convite para participar da pesquisa será encaminhado por e-mail (apêndice D) a diversos professores, informando o prazo de duas semanas para a coleta dos dados. Foram observadas as diretrizes da Lei Geral de Proteção de Dados (LGPD). Nenhuma informação que pudesse identificar individualmente os professores participantes foi coletada, assegurando o anonimato e a privacidade dos envolvidos. O consentimento dos participantes foi

obtido por meio do Termo de Consentimento Livre e Esclarecido (TCLE), disponibilizado no início do questionário, garantindo que todos estivessem plenamente informados sobre os objetivos, procedimentos e direitos relacionados à pesquisa.

Na avaliação automática, os próprios LLMs por meio de um *Prompt que pode ser encontrado no quadro 4*, serão utilizados para avaliar as questões de forma cruzada. Um LLMs gerará as questões, que serão subsequentemente analisadas pelos demais LLMs participantes junto com as questões do professor, com base nos critérios pré-definidos (Cobertura, Gramaticalidade, Utilidade, Relevância e Capacidade de resposta) atribuindo pontuações em uma escala pré-determinada (1 a 5) para cada critério, como (Maity; Deroy; Sarkar, 2025) utilizou em seu experimento. Esse processo será repetido para as questões criadas com e sem *Meta-prompt*, permitindo uma análise comparativa do impacto dessa técnica na qualidade das questões.

Quadro 4 - Criação do Prompt para que os LLMs avaliem as questões.

Prompt: Instrução Principal: Você atuará como um especialista em avaliação educacional e Língua Portuguesa. Sua tarefa é analisar criticamente a questão fornecida abaixo, destinada a alunos do 7º ano do Ensino Fundamental, e avaliá-la rigorosamente com base em cinco critérios específicos. Para cada critério, você deve atribuir uma pontuação inteira de 1 a 5 e fornecer uma justificativa textual concisa para sua pontuação.

Questão a ser Avaliada:

[QUESTÃO COMPLETO GERADA]

Critérios de Avaliação e Escala de Pontuação:

As definições de cada critério e da escala devem ser consideradas antes de atribuir as pontuações.

Cobertura: Mede a extensão em que a pergunta gerada abrange os

elementos essenciais do contexto/conteúdo pretendido. Avalia se detalhes-chave que deveriam ser abordados estão presentes.

Escala:

- 1: Poucos ou nenhum elemento relevante coberto.
- 2: Alguns elementos cobertos, mas detalhes importantes ausentes.
- 3: Vários elementos importantes cobertos, mas alguns detalhes negligenciados.
- 4: Cobertura ampla dos elementos essenciais.
- 5: Cobertura abrangente de todos os elementos essenciais.

Gramaticalidade: Foca na aderência da pergunta às regras e convenções gramaticais padrão. Examina estrutura linguística, sintaxe e clareza.

Escala:

- 1: Numerosos erros; difícil de entender.
- 2: Vários erros; significado parcialmente obscuro.
- 3: Alguns erros, geralmente compreensíveis.
- 4: Erros menores; maior parte clara e correta.
- 5: Sem erros; completamente clara e correta.

Utilidade: Indica o potencial valor da pergunta para inclusão na avaliação de um professor. Avalia se contribui efetivamente para a aprendizagem e avaliação, mesmo que não seja totalmente respondível pelo contexto.

Escala:

- 1: Não útil; nenhuma contribuição para a aprendizagem.

- 2: Utilidade limitada; impacto mínimo na avaliação.

- 3: Moderadamente útil; contribui para a aprendizagem, embora um pouco previsível.

- 4: Útil; melhora a avaliação e a compreensão do estudante.

- 5: Altamente útil; aumenta significativamente a aprendizagem e os resultados da avaliação.

Relevância: Avalia o alinhamento semântico entre a pergunta gerada e o contexto/conteúdo de entrada correspondente. Considera se mantém o foco nos conceitos-chave e evita ideias não relacionadas.

Escala:

- 1: Não relevante; sem conexão com o contexto.

- 2: Pouco relevante; conexão fraca com o contexto.

- 3: Moderadamente relevante; algum foco no contexto.

- 4: Relevante; alinhada de perto com o contexto.

- 5: Altamente relevante; aborda diretamente o contexto sem ideias não relacionadas.

Capacidade de Resposta: Mede se a pergunta pode ser respondida com base no contexto fornecido ou no conhecimento esperado. Não exige encontrar uma porção exata no contexto, mas sim se um estudante pode responder razoavelmente usando a lógica transmitida.

Escala:

- 1: Não respondível; sem base no contexto.

- 2: Difícil de responder; conexão fraca com o contexto.

•3: Moderadamente respondível; alguma lógica pode ser aplicada a partir do contexto.

•4: Quase respondível; base clara no contexto.

•5: Totalmente respondível; diretamente suportada pelo contexto.

Tarefa: Para a questão fornecida acima, forneça sua avaliação para CADA UM dos 5 critérios (Cobertura, Gramaticalidade, Utilidade, Relevância, Capacidade de Resposta). Para cada critério, você deve indicar:

•A pontuação (um número inteiro de 1 a 5).

•Uma justificativa textual breve (1-2 frases) explicando o porquê da pontuação atribuída.

Formato da Resposta: Por favor, formate sua resposta final estritamente como uma tabela com três colunas: Critério, Pontuação e Justificativa. Inclua uma linha para cada um dos cinco critérios avaliados.

Fonte: Elaborada pela autora (2025).

Essa etapa tem como objetivo estabelecer um padrão de referência para comparar as avaliações realizadas pelas duas abordagens. A preparação das questões envolveu a elaboração pelo professor especialista (Tabela B.1, Apêndice B) e a geração por LLMs, utilizando tanto *Prompt* (Tabela B.2, Apêndice B) quanto um *Meta-prompt* (Tabela B.4, Apêndice B). O fluxo geral desse processo está esquematizado na Figura 5.

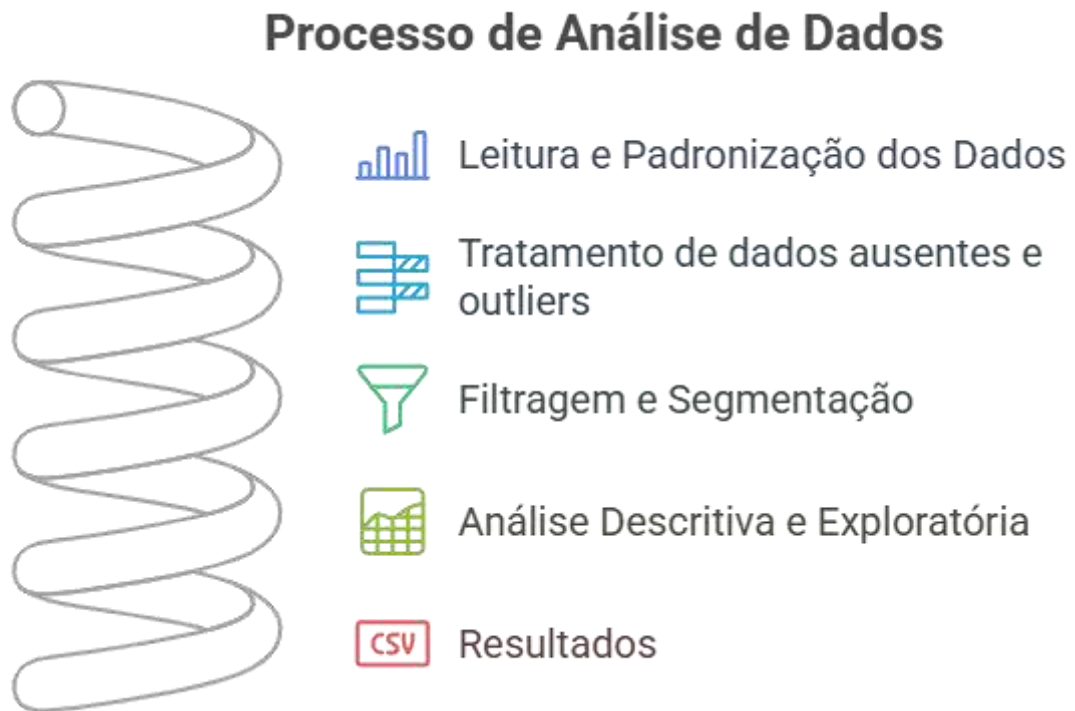


Figura 5 - Fluxograma do processo analítico quantitativo.

Fonte: Elaborada pela autora (2025).

- **Leitura e Padronização dos Dados:** Para a avaliação automatizada, os arquivos contêm identificadores das questões, o LLM gerador, o LLM avaliador, o tipo de *Prompt* utilizado, os critérios avaliados e as pontuações atribuídas. Os nomes dos identificadores serão padronizados, incluindo capitalização e remoção de espaços, para garantir a uniformidade dos agrupamentos e facilitar a análise.
- **Dados ausentes e *outliers*:** O questionário foi configurado com itens obrigatórios, assegurando a completude das respostas e aderindo às orientações éticas. Nas avaliações automáticas realizadas pelos LLMs, não foram verificados campos faltantes, uma vez que os *Prompts* foram estruturados para contemplar obrigatoriamente os cinco critérios e suas respectivas pontuações. A identificação de outliers será conduzida por inspeção gráfica e pelo critério do intervalo interquartil (IQR), garantindo a robustez da análise.
- **Filtragem e Segmentação:** Será aplicado um filtro nas avaliações automatizadas, denominadas “cruzadas”, considerando

exclusivamente as linhas em que tanto o gerador quanto o avaliador correspondem a LLMs. Essa abordagem exclui qualquer interação com o professor, permitindo investigar de modo puramente automatizado os padrões de avaliação e o rigor entre os modelos.

- **Análise Descritiva e Exploratória:** Serão calculadas estatísticas descritivas, como média, desvio padrão e contagem de avaliações, cruzando cada critério entre os dois tipos de *Prompt*. Adicionalmente, serão realizadas análises para classificar os LLMs por critério e identificar os avaliadores mais rigorosos.
- **Resultados:** Serão exploradas correlações entre critérios e variação intra/interavaliador e detecção de *outliers*.

5 RESULTADOS E DISCUSSÃO

Neste capítulo, serão apresentados e discutidos os resultados obtidos no experimento. A análise dos resultados será organizada nas seguintes seções: 5.1 Resultados; 5.2 Discussão.

5.1 Resultados

A análise foi realizada de duas formas: (i) avaliação humana, realizada por professores de Língua Portuguesa por meio de um questionário e (ii) avaliação automática, conduzida por um esquema de avaliação cruzada entre os próprios LLMs.

Na avaliação humana, professores avaliaram as questões usando a escala *Likert* de 1 a 5 (1 = muito baixa; 5 = muito alta) para cinco critérios: Capacidade de resposta, Cobertura, Gramaticalidade, Relevância e Utilidade. Na avaliação automática, adotou-se um procedimento de avaliação cruzada, no qual cada LLM avaliou as questões produzidas pelos outros LLMs e as da professora seguindo os mesmos critérios e escala. Por “avaliação cruzada” entende-se, que cada LLM atua como avaliador das questões geradas pelos outros LLMs.

5.1.1 Análise Automática

A análise automática entre os LLMs é fundamental para avaliar a consistência das pontuações e investigar potenciais vieses quando um modelo avalia conteúdos produzidos por si mesmo ou por outros modelos. Cada modelo gerou questões de Língua Portuguesa do 7º ano sob duas estratégias, *Prompt* e *Meta-Prompt* e, em seguida, todos os modelos avaliaram todas as questões. Esta seção foca exclusivamente nos LLMs e na comparação entre tipos de *Prompt* como podem ser vistos na tabela 3 a seguir.

Tabela 3 - Análise cruzada com avaliadores LLMs agregados na escala 1–5, reportando Mediana, Q1 (25%), Q3 (75%) e n por Critério × Gerador × Tipo de Prompt; LLMs avaliados em Prompt e Meta-prompt, Professor apenas em Prompt.

Critério	Gerador	Tipo de Prompt	Mediana	Q1 (25%)	Q3 (75%)	n
Capacidade de resposta	ChatGPT	<i>Prompt</i>	4	4	5	9
Capacidade de resposta	ChatGPT	<i>Meta-prompt</i>	4	4	5	9
Capacidade de resposta	<i>DeepSeek</i>	<i>Prompt</i>	4	4	5	9
Capacidade de resposta	<i>DeepSeek</i>	<i>Meta-prompt</i>	4	4	5	9
Capacidade de resposta	<i>Mistral</i>	<i>Prompt</i>	4	4	5	9
Capacidade de resposta	<i>Mistral</i>	<i>Meta-prompt</i>	4	4	5	9
Capacidade de resposta	Professor	<i>Professor</i>	4	4	4,75	6
Cobertura	ChatGPT	<i>Prompt</i>	4	3	4	9
Cobertura	ChatGPT	<i>Meta-prompt</i>	4	3	4	9
Cobertura	<i>DeepSeek</i>	<i>Prompt</i>	3	3	3	9
Cobertura	<i>DeepSeek</i>	<i>Meta-prompt</i>	4	4	4	9
Cobertura	<i>Mistral</i>	<i>Prompt</i>	3	3	4	9
Cobertura	<i>Mistral</i>	<i>Meta-prompt</i>	3	3	3	9
Cobertura	Professor	<i>Professor</i>	3	3	3	6
Gramaticalidade	ChatGPT	<i>Prompt</i>	5	5	5	9
Gramaticalidade	ChatGPT	<i>Meta-prompt</i>	5	5	5	9
Gramaticalidade	<i>DeepSeek</i>	<i>Prompt</i>	4	4	5	9
Gramaticalidade	<i>DeepSeek</i>	<i>Meta-prompt</i>	5	5	5	9
Gramaticalidade	<i>Mistral</i>	<i>Prompt</i>	4	4	5	9
Gramaticalidade	<i>Mistral</i>	<i>Meta-prompt</i>	5	4	5	9
Gramaticalidade	Professor	<i>Professor</i>	4	4	4,75	6
Relevância	ChatGPT	<i>Prompt</i>	5	4	5	9
Relevância	ChatGPT	<i>Meta-prompt</i>	5	4	5	9
Relevância	<i>DeepSeek</i>	<i>Prompt</i>	4	3	4	9
Relevância	<i>DeepSeek</i>	<i>Meta-prompt</i>	5	5	5	9
Relevância	<i>Mistral</i>	<i>Prompt</i>	5	4	5	9
Relevância	<i>Mistral</i>	<i>Meta-prompt</i>	4	4	4	9
Relevância	Professor	<i>Professor</i>	4	4	4	6
Utilidade	ChatGPT	<i>Prompt</i>	4	3	4	9
Utilidade	ChatGPT	<i>Meta-prompt</i>	4	3	4	9
Utilidade	<i>DeepSeek</i>	<i>Prompt</i>	3	3	3	9
Utilidade	<i>DeepSeek</i>	<i>Meta-prompt</i>	4	4	4	9
Utilidade	<i>Mistral</i>	<i>Prompt</i>	4	3	4	9
Utilidade	<i>Mistral</i>	<i>Meta-prompt</i>	3	3	3	9
Utilidade	Professor	<i>Professor</i>	3	3	3	6

Notas: Mediana = valor central; Q1 e Q3 = 25º e 75º percentis; n = número de

julgamentos (avaliadores LLMs).

Fonte: Elaborada pela autora (2025).

A tabela 3 apresenta os resultados da análise cruzada com avaliadores LLMs agregados para cada critério — Capacidade de resposta, Cobertura, Gramaticalidade, Relevância e Utilidade — considerando como geradores os LLMs (*ChatGPT*, *DeepSeek* e *Mistral*) e o Professor. Para cada combinação Critério × Gerador × Tipo de *Prompt* (com *Prompt* e *Meta-prompt* para os LLMs e o Professor), são reproduzidos: a Mediana das notas na escala *Likert* 1–5; os *quartis* Q1 (25%) e Q3 (75%) e o *n*, o número de julgamentos válidos utilizados no cálculo por célula. A ordenação é feita por Critério, Gerador e Tipo de *Prompt*.

Os LLMs apresentam desempenho geral superior ao Professor em quatro dos cinco critérios e empatam em Capacidade de resposta. Em Gramaticalidade e Relevância, vários LLMs alcançam mediana = 5 (por exemplo, *ChatGPT* em *Prompt* e *Meta-Prompt*), enquanto o Professor permanece em 4. Em Cobertura e Utilidade, os LLMs ficam entre 3–4, superando o Professor (mediana 3 em ambos). Capacidade de resposta é o critério menos discriminativo: todas as medianas ficam em 4 (LLMs e Professor), embora os LLMs mostrem ter teto um pouco mais alto pelos *quartis* superiores. No conjunto, o *ChatGPT* é o modelo mais consistente: mantém 5 em Gramaticalidade e Relevância em *Prompt* e *Meta-Prompt* e 4 nos demais critérios; o *DeepSeek* também se destaca quando usa *Meta-Prompt*; o *Mistral* alterna picos (5 em Relevância com *Prompt* e 5 em Gramaticalidade com *Meta-Prompt*) com quedas (3 em Utilidade com *Meta-Prompt*). Assim, LLMs são melhores que Professor em desempenho central, Gramaticalidade e Relevância são os critérios com maiores notas, e o *ChatGPT* aparece como melhor LLM no geral pela combinação de patamar e estabilidade. Sobre o efeito do *Meta-Prompt*, os resultados variam por modelo. O *ChatGPT* praticamente não muda entre *Prompt* e *Meta-Prompt* (medianas 4–5), sugerindo robustez ao regime. O *DeepSeek* melhora com meta (chega a 5 em Gramaticalidade e Relevância e eleva Cobertura/Utilidade para 4), indicando ganho claro. O *Mistral* tem efeito misto: melhora em Gramaticalidade com *Meta-Prompt* (para 5), mas cai em Relevância (de 5 para 4) e Utilidade (de 4 para 3). Em termos práticos, o *Meta-Prompt* ajuda especialmente em Gramaticalidade e Cobertura, enquanto empata com *Prompt* em Capacidade de resposta, Relevância e Utilidade quando visto de forma agregada.

Uma possível explicação para as notas mais baixas do Professor decorre de dois fatores combinados: (i) o tamanho amostral menor — nas células do Professor há $n = 6$ julgamentos, contra $n = 9$ para os LLMs — o que reduz a estabilidade da mediana e diminui a chance de alcançar o valor máximo da escala; (ii) a geração mais limitada, pois o Professor aparece apenas em “*Prompt*”, enquanto os LLMs também foram avaliados em “*Meta-prompt*”, procedimento que elevou o desempenho de alguns modelos (como o *DeepSeek*).

5.1.2 Análise do perfil dos respondentes

Esta seção apresenta a análise das avaliações realizadas por 81 professores de Língua Portuguesa sobre as 20 questões geradas pelos LLMs e pela professora especialista, as questões incluíam produções dos três LLMs (*ChatGPT*, *Mistral* e *DeepSeek*) com ambos os tipos de *Prompt* e *Meta-Prompt*, além de duas questões elaboradas pela professora especialista. Diferentemente da análise automática apresentada na seção anterior, onde os próprios LLMs avaliaram as questões, aqui temos a perspectiva humana especializada, fundamental para validar a aplicabilidade pedagógica dos LLMs no contexto educacional real. O perfil dos professores foi analisado para verificar se características pessoais e profissionais ajudam ou não nos julgamentos sobre as questões. Consideramos variáveis como instituição de ensino, tempo de experiência docente, frequência de elaboração de questões, formação acadêmica, nível de estresse, satisfação profissional e experiência prévia com LLMs.

Conforme demonstrado na Tabela 4, caracterizamos o perfil dos 81 respondentes a partir de idade, titulação, tipo de instituição, tempo de docência, prática de elaboração de questões, tempo semanal dedicado a exercícios/avaliações, além de bem-estar (estresse, satisfação) e familiaridade com tecnologias (ferramentas de criação e LLMs). As variáveis categóricas são descritas por frequências e percentuais, e as ordinais (estresse/satisfação, escala 1–5) por mediana e quartis (Q1–Q3), destacando a dispersão via IQR.

Tabela 4 - Perfil dos professores: distribuição de respostas por variável (frequências e percentuais).

Variável (número da questão)	Categorias mais frequentes
Idade (Q1)	40–49 anos (32,1%); 30–39 anos (24,7%)
Pós-graduação (Q2)	Sim, Especialização (42,0%); Sim, Mestrado (27,2%)
Instituição (Q3)	Escola pública municipal (21,0%); Outra (19,8%)
Tempo de docência (Q4)	11–15 anos (28,4%); Mais de 20 anos (24,7%)
Frequência de elaborar questões (Q5)	Sempre (37,0%); Frequentemente (30,9%)
Tempo semanal para itens (Q6)	3–4 horas (37,0%); 1–2 horas (25,9%)
Estresse (Q7)	5 (29,6%); 4 (28,4%)
Satisfação profissional (Q8)	5 (37,0%); 4 (24,7%)
Uso prévio de ferramenta para criar questões (Q9)	Sim (77,8%); Não (22,2%)
Experiência prévia com LLMs (Q10)	Não tive experiência prévia (63,0%); Sim (8,6%)
Crença: LLMs são úteis? (Q11)	Sim (93,8%); Não (6,2%)

Nota: Cada linha traz as categorias mais frequentes e o número de respostas válidas da variável.

Fonte: Elaborada pela autora (2025).

A tabela 4 apresenta, para cada variável do perfil (idade, pós-graduação, tipo de instituição, tempo de docência, frequência de elaboração de questões, tempo semanal dedicado à criação de itens, uso prévio de ferramentas de autoria, experiência prévia com LLMs e crença na utilidade de LLMs), as categorias de resposta e suas contagens absolutas (n), acompanhadas dos percentuais calculados apenas sobre os casos válidos daquela variável. A última coluna informa o n válido da variável.

O grupo de respondentes é majoritariamente experiente e pós-graduado. As faixas 30 – 49 anos concentram mais da metade do total, e o tempo de docência maior que 11 anos predomina. Em titulação, sobressaem Especialização com quase metade dos respondentes. Há diversidade institucional, redes municipal/estadual, particular, Institutos Federais e colégios militares, sugerindo contextos avaliativos distintos. Na

prática de elaboração de itens, a maioria elabora frequentemente as próprias questões e dedica 3 – 4 horas semanais, sinal de familiaridade com critérios de qualidade e com o trabalho de construção de questões. Em tecnologia, a maioria já usou ferramentas de criação, mas não tem experiência prévia com LLMs; ainda assim, 93,8% acreditam que LLMs podem ser úteis, um dado de aceitabilidade elevado.

A tabela 5 resume as distribuições de Estresse e Satisfação por meio de n (casos válidos), Q1 (25%), Mediana (valor central) e Q3 (75%), além do IQR (Q3–Q1), que expressa a dispersão das respostas na faixa intermediária.

Tabela 5 - Bem-estar docente (variáveis ordinais): estresse e satisfação na escala Likert 1 – 5.

Variável	Q1 (25%)	Mediana	Q3 (75%)	IQR
Estresse (Q7)	3,0	4,0	5,0	2,0
Satisfação (Q8)	3,0	4,0	5,0	2,0

Fonte: Elaborada pela autora (2025).

No quesito bem-estar, tanto Estresse quanto Satisfação profissional têm mediana = 4 (escala 1–5), indicando docentes geralmente satisfeitos, embora sob pressão. Com base nos resultados apresentados na tabela 4 e 5 sobre o perfil, serão testadas algumas hipóteses usando as respostas das notas do questionário.

1. Professores com mais anos de docência tendem a dar notas iguais ou mais baixas para as questões geradas pelos LLMs.
2. Quem tem Mestrado/Doutorado tende a dar notas mais altas em Gramaticalidade e Relevância e a ser mais consistente.
3. Esperam-se diferenças entre escolas pública, privada e militar nas notas de Utilidade e Cobertura.
4. Quem elabora questões com mais frequência e dedica mais horas tende a ser mais crítico em Utilidade.
5. Quem acredita que LLMs são úteis tende a dar notas mais altas às questões geradas por LLM, mesmo sem experiência prévia.

6. Mais estresse tende a vir com notas mais baixas e/ou maior variação nas avaliações.

As comparações entre perfis docentes são apresentadas por estatísticas descritivas (n, mediana, Q1, Q3 e IQR) sem inferência estatística.

1. As medianas variam entre faixas de tempo de docência, mas não seguem um padrão crescente ou decrescente claro. Em alguns critérios, faixas intermediárias (6–10 anos) ou muito altas (>20 anos) apresentam patamares elevados, enquanto outras ficam no meio do caminho.
2. As medianas de Gramaticalidade e Relevância são semelhantes entre níveis de titulação (Especialização, Mestrado, Doutorado). A consistência (IQR) muda entre os níveis, mas sem um padrão definido em alguns casos, os níveis intermediários são os mais consistentes. Não se observa ganho sistemático com maior titulação.
3. As distribuições de notas por rede são muito próximas, com sobreposição de *quartis*, não se observam diferenças relevantes no conjunto.
4. Observa-se padrão oposto ao esperado: docentes que elaboram questões com maior frequência tendem a atribuir medianas mais altas em Utilidade. O efeito do tempo semanal dedicado é discreto e não aponta para maior rigor.
5. Há tendência de medianas mais altas entre quem acredita que LLMs são úteis, mais evidente entre docentes sem experiência prévia. As amostras são desiguais e há efeito teto, a leitura é descritiva.
6. Observa-se associação negativa entre nível de estresse e as notas: grupos com estresse mais alto apresentam medianas menores especialmente em Gramaticalidade, Relevância e Utilidade.

No conjunto, não há sinais de que senioridade, titulação ou rede produzam diferenças sistemáticas nas notas. Por outro lado, a crença na utilidade e prática frequente na elaboração de itens parecem elevar a percepção de Utilidade das respostas dos LLMs, enquanto maior estresse se relaciona a avaliações mais baixas. Esses achados ajudam a interpretar a heterogeneidade nas avaliações humanas e

sugerem pontos de atenção para formação docente, gestão do bem-estar e adoção pedagógica de IA. Para tabelas com análises detalhadas, veja o Apêndice E.

5.1.3 Análise da avaliação do Questionário

A descrição do perfil dos docentes fundamenta a etapa subsequente do estudo, a avaliação das 20 questões segundo cinco critérios, Capacidade de resposta, Cobertura, Gramaticalidade, Relevância e Utilidade em escala *Likert* 1–5. Inicialmente, apresentam-se os resumos descritivos globais por critério (Tabela 6), reportando n, mediana e quartis (Q1 e Q3). Na sequência, a Tabela 6 aprofunda a leitura ao estratificar por Gerador × Tipo (LLMs com *Prompt* e *Meta-prompt*; Professora com *Prompt*), preservando as mesmas estatísticas para comparabilidade. As diferenças de tamanho amostral por célula (243 para combinações LLM × tipo; 162 para Professora) decorrem do número de questões avaliadas em cada caso. Dado que medianas não são aditivas, pequenas diferenças entre o panorama global e o estratificado são esperadas, interpretações devem considerar n e quartis em cada estrato.

Tabela 6 — Estatísticas globais por critério (n, mediana, Q1 e Q3).

Critério	n	Mediana	Q1	Q3
Capacidade de resposta	1620	5,00	4,00	5,00
Cobertura	1620	4,00	4,00	5,00
Gramaticalidade	1620	5,00	4,00	5,00
Relevância	1620	4,00	4,00	5,00
Utilidade	1620	4,00	4,00	5,00

Nota: Escala: 1–5. n: número total de julgamentos por critério (todas as questões e avaliadores juntos). Q1/Q3: 25º/75º percentis. Na tabela agrega todas as questões

produzidas por todos os geradores (LLMs e professora), resultando em n = 1.620 por critério (≈ 81 docentes × 20 questões).

Fonte: Elaborada pela autora (2025).

A tabela 6 resume as notas por critério, as estatísticas das notas (1 – 5) por critério: Capacidade de resposta, Cobertura, Gramaticalidade, Relevância e Utilidade. Em todos os casos, n = 1.620 julgamentos. Observa-se mediana = 5,00 em Gramaticalidade e Capacidade de resposta (Q1 = 4,00; Q3 = 5,00), e mediana = 4,00 em Cobertura, Relevância e Utilidade (Q1 = 4,00; Q3 = 5,00). Assim, os *quartis* superiores alcançam 5,00 em todos os critérios, sugerindo presença de efeito teto.

O perfil global indica excelente forma textual com o critério de Gramaticalidade e boa Capacidade de resposta, ambos com mediana 5,00. Cobertura, Relevância e Utilidade mantêm mediana 4,00 com Q3 = 5,00, mostrando desempenho alto, porém com maior espaço para distinção entre itens. A combinação Q1 = 4,00 / Q3 = 5,00 em todos os critérios revela compressão no topo (efeito teto), reduzindo diferenças finas. Na tabela 7 podemos ver as estatísticas para de cada critério, *Prompt* e *Meta Prompt*.

Tabela 7 — Estatísticas por critério × gerador × tipo (n, mediana, Q1 e Q3).

Critério	Gerador	Tipo de Prompt	Mediana	Q1	Q3
Capacidade de resposta	ChatGPT	<i>Prompt</i>	4,00	3,00	5,00
Capacidade de resposta	ChatGPT	<i>Meta-prompt</i>	5,00	4,00	5,00
Capacidade de resposta	DeepSeek	<i>Prompt</i>	5,00	4,00	5,00
Capacidade de resposta	DeepSeek	<i>Meta-prompt</i>	5,00	4,00	5,00
Capacidade de resposta	Mistral	<i>Prompt</i>	5,00	4,00	5,00

Capacidade de resposta	<i>Mistral</i>	<i>Meta-prompt</i>	5,00	4,00	5,00
Capacidade de resposta	Professor	<i>Professor</i>	5,00	4,00	5,00
Cobertura	ChatGPT	<i>Prompt</i>	4,00	4,00	5,00
Cobertura	ChatGPT	<i>Meta-prompt</i>	4,00	4,00	5,00
Cobertura	<i>DeepSeek</i>	<i>Prompt</i>	4,00	4,00	5,00
Cobertura	<i>DeepSeek</i>	<i>Meta-prompt</i>	4,00	4,00	5,00
Cobertura	<i>Mistral</i>	<i>Prompt</i>	4,00	3,00	5,00
Cobertura	<i>Mistral</i>	<i>Meta-prompt</i>	4,00	4,00	5,00
Cobertura	Professor	<i>Professor</i>	4,50	4,00	5,00
Gramaticalidade	ChatGPT	<i>Prompt</i>	5,00	4,00	5,00
Gramaticalidade	ChatGPT	<i>Meta-prompt</i>	5,00	4,00	5,00
Gramaticalidade	<i>DeepSeek</i>	<i>Prompt</i>	5,00	4,00	5,00
Gramaticalidade	<i>DeepSeek</i>	<i>Meta-prompt</i>	5,00	4,00	5,00
Gramaticalidade	<i>Mistral</i>	<i>Prompt</i>	5,00	4,00	5,00

Gramaticalidade	<i>Mistral</i>	<i>Meta-prompt</i>	5,00	4,00	5,00
Gramaticalidade	Professor	<i>Professor</i>	5,00	4,00	5,00
Relevância	ChatGPT	<i>Prompt</i>	4,00	4,00	5,00
Relevância	ChatGPT	<i>Meta-prompt</i>	5,00	4,00	5,00
Relevância	<i>DeepSeek</i>	<i>Prompt</i>	4,00	4,00	5,00
Relevância	<i>DeepSeek</i>	<i>Meta-prompt</i>	4,00	4,00	5,00
Relevância	<i>Mistral</i>	<i>Prompt</i>	4,00	4,00	5,00
Relevância	<i>Mistral</i>	<i>Meta-prompt</i>	4,00	4,00	5,00
Relevância	Professor	<i>Professor</i>	5,00	4,00	5,00
Utilidade	ChatGPT	<i>Prompt</i>	4,00	3,00	5,00
Utilidade	ChatGPT	<i>Meta-prompt</i>	4,00	3,50	5,00
Utilidade	<i>DeepSeek</i>	<i>Prompt</i>	4,00	4,00	5,00
Utilidade	<i>DeepSeek</i>	<i>Meta-prompt</i>	4,00	4,00	5,00
Utilidade	<i>Mistral</i>	<i>Prompt</i>	4,00	3,00	5,00

Utilidade	<i>Mistral</i>	<i>Meta-prompt</i>	4,00	4,00	5,00
Utilidade	Professor	Professor	4,00	4,00	5,00

Nota: Escala: 1–5. n: julgamentos por combinação (LLM × tipo geralmente n = 243; Professor n = 162, pois há menos itens) para Professor, n = 162 porque há 2 questões avaliadas (2×81). Geradores: ChatGPT, *Mistral*, *DeepSeek*, Professora. Tipo: *Prompt* e *Meta-prompt* (para LLMs); para a Professora, “Professor”. A tabela mostra, para cada critério, as estatísticas separadas por gerador e tipo, permitindo comparar desempenhos.

Fonte: Elaborada pela autora (2025).

A tabela 7 apresenta, para cada critério (Capacidade de resposta, Cobertura, Gramaticalidade, Relevância e Utilidade), as estatísticas das notas (n, mediana, Q1 e Q3) por gerador (*ChatGPT*, *Mistral*, *DeepSeek*, Professor) e tipo (*Prompt*, *Meta-Prompt*; no caso do Professor, “Professor”). O tamanho amostral por célula é n = 243 para combinações LLM × tipo e n = 162 para Professor. Em Capacidade de resposta, a mediana = 5,00 em todas as combinações, exceto ChatGPT – Prompt (4,00); com *Meta-prompt*, *ChatGPT* também atinge 5,00 (*quartis*, em geral, Q1 = 4,00 e Q3 = 5,00). Em Cobertura, os LLMs ficam em 4,00 (com *Mistral–Prompt* Q1 = 3,00), enquanto o Professor registra 4,50 (Q1 = 4,00; Q3 = 5,00). Em Gramaticalidade, todos (LLMs e Professor) têm mediana = 5,00 (Q1 = 4,00; Q3 = 5,00). Em Relevância, *ChatGPT–Meta-prompt* e Professor alcançam 5,00 (demais LLMs em 4,00, com Q1 = 4,00; Q3 = 5,00). Em Utilidade, há mediana = 4,00 para todas as combinações; os *quartis* diferenciam os perfis: *ChatGPT – Prompt* e *Mistral – Prompt* têm Q1 = 3,00, *ChatGPT – Meta-prompt* Q1 = 3,50, e *DeepSeek (Prompt/Meta)* e Professor Q1 = 4,00; em todas as células Q3 = 5,00, indicando efeito teto frequente.

A Tabela 7 organiza os resultados por Critério × Gerador × Tipo, permitindo enxergar diferenças que a visão global não mostra. As medianas são, em geral, altas, e os *quartis* ajudam a separar melhor os perfis: com Q3 = 5,0 em quase todas as células, o que muda entre combinações é sobretudo o Q1. Dentro dos modelos, observa-se sensibilidade ao tipo de *Prompt*. No *ChatGPT*, o *Meta-prompt* melhora a

Capacidade de resposta e a Relevância (mediana 5,0) e encurta a cauda inferior em Utilidade (Q1 sobe de 3,0 para 3,5), o que indica menos ocorrências de notas baixas. No *Mistral*, o *Meta-prompt* aumenta o Q1 em Cobertura e Utilidade (de 3,0 para 4,0), sinalizando ganho de robustez sob os mesmos critérios. Já o *DeepSeek* se mantém estável entre *Prompt* e *Meta-prompt*, preservando medianas e Q1 = 4,0 nos critérios mais sujeitos a variação, o que sugere consistência independentemente do tipo. Em paralelo, o Professor apresenta vantagem sutil em Cobertura (mediana 4,5 e Q1 = 4,0), enquanto em Gramaticalidade e Capacidade de resposta há convergência no topo entre todos os geradores (mediana 5,0; Q1 = 4,0), refletindo qualidade textual elevada e pouco espaço para discriminação pela mediana.

Na prática, as diferenças que realmente fazem a diferença aparecem nas células com Q1=3,0, como por exemplo no *ChatGPT – Prompt* e *Mistral – Prompt* em Utilidade, e no *Mistral – Prompt* em Cobertura. Nesse caso há maior variabilidade entre avaliadores e risco de quedas locais de qualidade percebida. O *Meta-prompt* contribui mais reduzindo a cauda inferior do que elevando o topo, melhorando a consistência das notas. Em termos de seleção e refino de itens, vale priorizar as combinações e questões em que o Q1 é baixo e/ou o IQR é alto ($\geq 1,0$), pois são justamente os casos com divergência substantiva entre docentes e, portanto, maior potencial de melhora via ajustes de *Prompt* e revisão de conteúdo. Por fim, como as medianas não se agregam linearmente, pequenas diferenças entre a visão global e a estratificada são naturais.

5.2 Discussão

A presente seção dedica-se à interpretação e à análise crítica dos resultados na seção 5.1, contextualizando-os no âmbito dos objetivos que nortearam este estudo. Os achados revelam-se particularmente significativos por contrariarem as expectativas iniciais, oferecendo uma perspectiva nuançada sobre o papel e a eficácia das estratégias de *Prompt* na geração de conteúdo educacional por LLMs.

Um dos resultados centrais da análise estatística, conforme evidenciado na seção 5.1, é a ausência de diferenças significativas entre a qualidade das questões geradas por meio de *Prompt* e aquelas originadas com o *Meta-Prompt*, em todos os critérios avaliados (Cobertura, Gramaticalidade, Relevância, Utilidade e Capacidade

de Resposta). Esta observação contesta diretamente a expectativa de uma superioridade do *Meta-Prompt* no contexto analisado. Uma interpretação plausível para este fenômeno está na crescente sofisticação dos LLMs atuais. É possível que modelos como o *ChatGPT*, *Mistral* e *Deepseek* utilizados no experimento já possuam uma capacidade intrínseca de inferir a intenção e o contexto pedagógico a partir de instruções diretas e menos elaboradas. Para a tarefa específica de criar questões para o 7º ano do Ensino Fundamental, a complexidade pode não ter sido suficientemente elevada para que os benefícios de uma instrução mais granular e estruturada, como a do *Meta-Prompt*, se manifestassem de forma estatisticamente relevante. Para este domínio, os *Prompts* podem já ter fornecido informação suficiente para que os modelos operassem perto de seu desempenho ótimo, tornando o detalhamento adicional do *Meta-Prompt* redundante.

O segundo achado de maior relevância é o desempenho comparável entre as questões elaboradas por uma professora especialista e aquelas geradas pelos LLMs, independentemente da técnica de *Prompt* utilizada. A paridade de qualidade em critérios como Cobertura, Gramaticalidade, Relevância e Utilidade, observada tanto na análise automática (realizada pelos próprios LLMs de forma cruzada) quanto na análise humana (efetuada por professores especialistas), sugere que as ferramentas de IA atingiram um grau de maturidade que as torna recursos viáveis e eficazes para auxiliar os docentes. A concordância entre as duas abordagens de avaliação reforça a robustez desses achados.

Adicionalmente, a análise do perfil dos professores, considerando variáveis como instituição de atuação, tempo de experiência docente, frequência de elaboração de questões, formação acadêmica, nível de estresse, satisfação profissional e experiência prévia com LLMs, revelou que esses fatores não alteraram significativamente a percepção geral sobre a qualidade das questões geradas pelos LLMs em comparação com as elaboradas por humanos. Embora as percepções individuais possam variar, a tendência agregada dos avaliadores humanos corroborou os resultados da análise automática, indicando uma aceitação e reconhecimento da qualidade das questões produzidas pelos LLMs, independentemente das características demográficas ou profissionais dos docentes.

Tal constatação não aponta para uma substituição do professor, mas sim para

uma redefinição de seu papel, que pode evoluir de criador para curador de conteúdo. Os LLMs podem funcionar como assistentes eficientes, gerando uma base de questões de alta qualidade que o professor pode, adaptar e contextualizar à realidade de sua sala de aula. Isso representa um potencial significativo para a otimização do tempo do educador, permitindo que ele se concentre em atividades de maior valor agregado, como o acompanhamento individualizado dos alunos e o planejamento de estratégias de ensino inovadoras.

Ao reunir os resultados da análise automática e humana, surge uma implicação ampla: se os LLMs conseguem produzir material de qualidade comparável à de um especialista a partir de comandos simples, a barreira para a adoção de IA na educação é reduzida. Professores podem extrair grande valor dessas ferramentas no processo de ensino-aprendizagem. Contudo, é importante ressaltar que estes achados são válidos no contexto deste estudo. A generalização para outras disciplinas, níveis de ensino ou modelos de linguagem deve ser feita com cautela. Embora a estratégia de *Meta-Prompt* não tenha se provado superior, a prontidão dos LLMs para servirem como ferramentas de apoio e acessíveis na criação de conteúdo educacional.

5.3 Confirmação/Refutação das Hipóteses

As hipóteses foram avaliadas à luz dos resultados quantitativos obtidos nas seções anteriores, considerando tanto a análise cruzada entre LLMs quanto às avaliações realizadas pelos professores especialistas (unidade analítica: mediana por questão, escala 1–5).

H1: A aplicação de *Meta-Prompt* em LLMs eleva a qualidade das questões geradas em comparação com o uso de *Prompts*, refletindo-se em uma pontuação superior nos critérios definidos.

A hipótese de que o uso de *Meta-Prompt* em LLMs elevaria a qualidade das questões geradas em relação aos *Prompts* simples não foi confirmada. As análises não identificaram diferenças estatisticamente significativas entre os dois tipos de instrução em nenhum dos cinco critérios avaliados — Cobertura, Gramaticalidade, Relevância, Utilidade e Capacidade de Resposta. Observou-se apenas uma melhora pontual na consistência dos resultados, com aumento de valores mínimos (Q1) em alguns casos: no ChatGPT, a mediana de Capacidade de Resposta

alcançou 5,0 e o Q1 de Utilidade subiu de 3,0 para 3,5; no *Mistral*, houve elevação do Q1 em Cobertura e Utilidade de 3,0 para 4,0. Ainda assim, as medianas permaneceram inalteradas e a recorrência de valores máximos (efeito teto) impediu a caracterização de uma vantagem global do *Meta-Prompt*. Portanto, o *Meta-Prompt* não se mostrou superior aos prompts convencionais, apesar de manter a coerência e a adequação pedagógica das questões geradas.

H2: LLMs, quando adequadamente instruídos, são capazes de gerar questões de Língua Portuguesa com qualidade pedagógica equivalente à de um professor, não havendo diferença estatisticamente significativa entre as avaliações de ambos os grupos.

A segunda hipótese, que propunha que os LLMs, quando adequadamente instruídos, poderiam gerar questões de Língua Portuguesa com qualidade equivalente às de um professor, foi confirmada. As questões produzidas pelos modelos apresentaram desempenho comparável às da professora tanto na avaliação humana quanto na automática. Em Gramaticalidade e Capacidade de Resposta, todas as medianas foram 5,0. A professora obteve leve vantagem em Cobertura (4,5 frente a 4,0 dos LLMs), enquanto a Relevância atingiu 5,0 tanto para a professora quanto para o ChatGPT com *Meta-Prompt* a Utilidade manteve mediana de 4,0 em todos os casos. Variáveis de perfil docente não alteraram essas conclusões. Esses resultados indicam que os LLMs são capazes de produzir questões de qualidade pedagógica equivalente à humana, quando orientados de forma adequada.

5.4 Resposta às Questões de Pesquisa

Com base nos resultados obtidos, é possível responder às questões de pesquisa propostas neste estudo:

QP1: A utilização de *Meta-Prompt* em LLMs resulta em questões de Língua Portuguesa com qualidade pedagógica significativamente superior às aquelas geradas por meio de prompts simples, na percepção de docentes?

Os resultados demonstram que não houve diferença significativa entre os dois tipos de instrução. O *Meta-Prompt* apresentou leve melhora na consistência das

notas mais baixas, mas não gerou ganhos significativos nas medianas dos critérios avaliados. Embora ofereça estrutura mais detalhada e maior controle sobre o processo de geração, não foi percebida melhoria clara na qualidade pedagógica das questões. Assim, QP1 foi respondida negativamente: o *Meta-Prompt* não superou o *Prompt* na percepção dos docentes.

QP2: Questões de Língua Portuguesa geradas por LLMs, com ou sem *Meta-Prompt*, possuem qualidade comparável àquelas elaboradas por um professor especialista, considerando critérios como relevância, utilidade e adequação curricular?

As análises indicaram paridade entre as questões geradas pelos LLMs e as da professora em todos os critérios avaliados. A docente apresentou pequena vantagem em Cobertura, mas os LLMs igualaram o desempenho em Gramaticalidade, Relevância, Utilidade e Capacidade de Resposta, todas com medianas altas e próximas. A concordância entre as avaliações humanas e automáticas reforça que as produções dos LLMs são comparáveis às humanas, demonstrando sua viabilidade como ferramentas de apoio docente. Dessa forma, QP2 foi respondida positivamente, confirmando que os modelos de linguagem podem auxiliar com eficácia na elaboração de exercícios de Língua Portuguesa.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Após a apresentação e discussão dos resultados obtidos, esta seção final da dissertação dedica-se a consolidar as principais conclusões do estudo, bem como a delinear suas limitações e as perspectivas para pesquisas futuras.

6.1 Limitações do Estudo

Este estudo, embora tenha fornecido insights valiosos sobre a aplicação de LLMs na geração de questões educacionais, apresenta algumas limitações inerentes ao seu escopo. Primeiramente, a pesquisa concentrou-se exclusivamente na disciplina de Língua Portuguesa para alunos do 7º ano do Ensino Fundamental. Essa especificidade, embora crucial para a profundidade da análise, restringe a generalização dos resultados para outras áreas do conhecimento, níveis de ensino ou faixas etárias. A complexidade intrínseca de diferentes disciplinas e as nuances pedagógicas associadas a cada etapa educacional podem influenciar significativamente o desempenho dos LLMs e a eficácia das estratégias de *Prompt*.

A avaliação da qualidade das questões foi realizada dentro de um contexto e parametrização específicos. A escolha dos LLMs, as configurações padrão utilizadas e os critérios de avaliação (Cobertura, Gramaticalidade, Relevância, Utilidade e Capacidade de Resposta) foram determinantes para os achados. Variações nesses parâmetros, como a utilização de outros modelos de linguagem, ajustes finos ou a aplicação de diferentes métricas de avaliação, poderiam alterar os resultados observados. A ausência de diferenças significativas entre prompts simples e meta-prompts, por exemplo, pode estar ligada ao nível de complexidade da tarefa proposta, que talvez não tenha exigido o potencial máximo das estratégias mais elaboradas.

Por fim, a amostra de professores envolvidos na avaliação humana, embora composta por especialistas, pode não representar a totalidade da diversidade de perfis docentes, experiências e percepções sobre o uso de tecnologias educacionais. A subjetividade inerente à avaliação humana, mesmo quando guiada por critérios claros, constitui um fator que, embora mitigado, não pode ser completamente eliminado.

6.2 Trabalhos Futuros

Com base nas limitações identificadas e nos resultados obtidos, diversas pesquisas futuras podem ser exploradas para aprofundar a compreensão sobre o uso de LLMs na educação:

1. Expansão para Outras Disciplinas e Níveis de Ensino: Investigar a aplicabilidade e a eficácia das estratégias de *Prompt* na geração de questões para outras disciplinas e em diferentes níveis de ensino. Isso permitiria verificar se a paridade de desempenho entre LLMs e professores se mantém em contextos de maior complexidade conceitual ou exigência pedagógica.
2. Estudo da Percepção e Aceitação Docente: Aprofundar a investigação sobre a percepção dos professores em relação às questões geradas por LLMs, explorando fatores como a confiança na ferramenta, a facilidade de integração no fluxo de trabalho e o impacto na carga de trabalho. Isso pode incluir estudos qualitativos, como entrevistas e grupos focais, para capturar nuances que avaliações quantitativas não abordam.
3. Desenvolvimento de Estratégias de *Prompt* Adaptativas: Pesquisar e desenvolver estratégias de *Meta-Prompt* mais sofisticadas e adaptativas, que possam ser ajustadas dinamicamente com base no contexto, na disciplina e no nível de complexidade desejado. Isso poderia incluir a incorporação de *feedback* iterativo dos professores para otimizar a geração de questões.
4. Avaliação do Impacto no Aprendizado dos Alunos: Um passo crucial seria investigar o impacto direto das questões geradas por LLMs no desempenho e no engajamento dos alunos. Estudos que comparem o progresso de estudantes expostos a questões geradas por IA versus questões elaboradas por humanos poderiam fornecer evidências empíricas sobre o valor pedagógico dessas ferramentas.

REFERÊNCIA

ABRAMOVAY; MIRIAM; RUA; MARIA DAS GRAÇAS. **Violências nas escolas: versão resumida - UNESCO Digital Library**. [no place]: [no publisher], 2003. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000133967_por>. Acesso em: 08/jul./24.

AL FARABY, Said; ADIWIJAYA, Adiwijaya; ROMADHONY, Ade. **Review on Neural Question Generation for Education Purposes**. In: *International Journal of Artificial Intelligence in Education*, v. 34, n° 3, p. 1008–1045, 2024. ISSN: 1560-4306, DOI: 10.1007/s40593-023-00374-x.

AMATRIAIN, Xavier; SANKAR, Ananth; BING, Jie; BODIGUTLA, Praveen Kumar; HAZEN, Timothy J.; KAZI, Michael. **Transformer models: an introduction and catalog**. arXiv, 2024. Disponível em: <<http://arxiv.org/abs/2302.07730>>. Acesso em: 22/jun./24

ANDERSON, Lorin W.; KRATHWOHL, David R. **A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives: complete edition**. [no place]: Addison Wesley Longman, Inc., 2001. ISBN: 978-0-321-08405-7. Disponível em: <<https://eduq.info/xmlui/handle/11515/18824>>. Acesso em: 25/jun./25.

BILLING, David. **Teaching for transfer of core/key skills in higher education: Cognitive skills**. In: *Higher Education*, v. 53, n° 4, p. 483–516, 2007. ISSN: 1573-174X, DOI: 10.1007/s10734-005-5628-5.

BROWN, Tom B.; MANN, Benjamin; RYDER, Nick; SUBBIAH, Melanie; KAPLAN, Jared; DHARIWAL, Prafulla; NEELAKANTAN, Arvind; SHYAM, Pranav; SASTRY, Girish; ASKELL, Amanda; AGARWAL, Sandhini; HERBERT-VOSS, Ariel; KRUEGER, Gretchen; HENIGHAN, Tom; CHILD, Rewon; RAMESH, Aditya; ZIEGLER, Daniel M.; WU, Jeffrey; WINTER, Clemens; HESSE, Christopher; CHEN, Mark; SIGLER, Eric; LITWIN, Mateusz; GRAY, Scott; CHESS, Benjamin; CLARK, Jack; BERNER, Christopher; MCCANDLISH, Sam; RADFORD, Alec; SUTSKEVER, Ilya; AMODEI, Dario. **Language models are few-shot learners**. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20*, 2020, Red Hook, NY, USA. Red Hook, NY, USA: Curran Associates Inc. 2020. p. 1877–1901. Acesso em: 22/abr./25.

CARDOSO, Fábio Santos; PEREIRA, Natália da Silva; BRAGGION, Rodrigo César; CHAVES, Paloma; ANDRIOLI, Mary. **O uso da Inteligência Artificial na Educação**

e seus benefícios: uma revisão exploratória e bibliográfica. *In: Revista Ciência em Evidência*, v. 4, nº FC, p. e023002–e023002, 2023. ISSN: 2763-5457, DOI: 10.47734/rce.v4iFC.2332.

CHAN, Winston; AN, Aijun; DAVOUDI, Heidar. **A Case Study on ChatGPT Question Generation.** *In: 2023 IEEE International Conference on Big Data (BigData)*, 2023. [no place]: [no publisher]. 2023. p. 1647–1656. Disponível em: <<https://ieeexplore.ieee.org/document/10386520>>. Acesso em: 09/maio/25.

CHOWDHURY, Gobinda G. **Natural language processing.** *In: Annual Review of Information Science and Technology*, v. 37, nº 1, p. 51–89, 2003. ISSN: 1550-8382, DOI: 10.1002/aris.1440370103.

CRONBACH, Lee J. **Course Improvement through Evaluation.** *In: Teachers College Record: The Voice of Scholarship in Education*, v. 64, nº 8, p. 1–13, 1963. ISSN: 0161-4681, 1467-9620, DOI: 10.1177/016146816306400802.

ELKINS, Sabina; KOCHMAR, Ekaterina; CHEUNG, Jackie C. K.; SERBAN, Iulian. **How Teachers Can Use Large Language Models and Bloom's Taxonomy to Create Educational Quizzes.** *In: Proceedings of the AAAI Conference on Artificial Intelligence*, v. 38, nº 21, p. 23084–23091, 2024. ISSN: 2374-3468, DOI: 10.1609/aaai.v38i21.30353.

ELKINS, Sabina; KOCHMAR, Ekaterina; SERBAN, Iulian; CHEUNG, Jackie C. K. **How Useful Are Educational Questions Generated by Large Language Models?** *In: Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, 2023, Cham. Cham: Springer Nature Switzerland. 2023. p. 536–542.

EWELL, Peter T. **Assessment and accountability in America today: Background and context.** *In: New Directions for Institutional Research*, v. 2008, nº S1, p. 7–17, 2008. ISSN: 1536-075X, DOI: 10.1002/ir.258.

FREIRE, Paulo. **Pedagogia do Oprimido.** [no place]: [no publisher], 1970.

FREIRE, Paulo. **Pedagogia da Autonomia.** [no place]: [no publisher], 1996.

GALLEGOS, Isabel O.; ROSSI, Ryan A.; BARROW, Joe; TANJIM, Md Mehrab; KIM, Sungchul; DERNONCOURT, Franck; YU, Tong; ZHANG, Ruiyi; AHMED, Nesreen K.

Bias and Fairness in Large Language Models: A Survey. *In: Computational Linguistics*, v. 50, nº 3, p. 1097–1179, 2024. ISSN: 0891-2017, DOI: 10.1162/coli_a_00524.

GOMES, Dennis Dos Santos. **Inteligência Artificial: Conceitos e Aplicações.** *In: Inteligência Artificial*, 2010.

HE, Xuehai; YANG, Diji; FENG, Weixi; FU, Tsu-Jui; AKULA, Arjun; JAMPANI, Varun; NARAYANA, Pradyumna; BASU, Sugato; WANG, William Yang; WANG, Xin. **CPL: Counterfactual Prompt Learning for Vision and Language Models.** *In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, Abu Dhabi, United Arab Emirates. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. 2022. p. 3407–3418. Disponível em: <<https://aclanthology.org/2022.emnlp-main.224/>>. Acesso em: 10/nov./25.

HOLMES, Wayne; PORAYSKA-POMSTA, Kaska; HOLSTEIN, Ken; SUTHERLAND, Emma; BAKER, Toby; SHUM, Simon Buckingham; SANTOS, Olga C.; RODRIGO, Mercedes T.; CUKUROVA, Mutlu; BITTENCOURT, Ig Ibert; KOEDINGER, Kenneth R. **Ethics of AI in Education: Towards a Community-Wide Framework.** *In: International Journal of Artificial Intelligence in Education*, v. 32, nº 3, p. 504–526, 2022. ISSN: 1560-4306, DOI: 10.1007/s40593-021-00239-1.

HOWARD, Tyrone C.; ALEMAN, Glenda R. **Teacher capacity for diverse learners: What do teachers need to know?** *In: Handbook of Research on Teacher Education*. 3 ed. [no place]: Routledge, 2008. 18 p. ISBN: 978-0-203-93869-0.

IONESCU, Valeriu Manuel; ENESCU, Madalin Ciprian. **Using ChatGPT for Generating and Evaluating Online Tests.** *In: 2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 2023, Bucharest, Romania. Bucharest, Romania: IEEE. 2023. p. 1–6. Disponível em: <<https://ieeexplore.ieee.org/document/10193995/>>. Acesso em: 04/fev./25.

JUNQUEIRA, Julia da Rocha; CORRÊA, Ulisses Brisolara; FREITAS, Larissa. **Transformer Models for Brazilian Portuguese Question Generation: An Experimental Study.** *In: The International FLAIRS Conference Proceedings*, v. 37, 2024. ISSN: 2334-0762, DOI: 10.32473/flairs.37.1.135334. Disponível em: <<https://journals.flvc.org/FLAIRS/article/view/135334>>. Acesso em: 16/fev./25.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and**

Speech Recognition. 3 ed. ed. [no place]: [no publisher], 2023. Disponível em: <https://www.researchgate.net/publication/200111340_Speech_and_Language_Processing_An_Introduction_to_Natural_Language_Processing_Computational_Linguistics_and_Speech_Recognition>. Acesso em: 25/jun./25.

KLIEBARD, Herbert M. **Os princípios de Tyler**. In: *Currículo sem fronteiras*, v. 11, nº 2, p. 23–35, 2011.

KOJIMA, Takeshi; GU, Shixiang Shane; REID, Machel; MATSUO, Yutaka; IWASAWA, Yusuke. **Large Language Models are Zero-Shot Reasoners**. arXiv, 2023. Disponível em: <<http://arxiv.org/abs/2205.11916>>. Acesso em: 25/jun./25

KONG, Aobo; ZHAO, Shiwan; CHEN, Hao; LI, Qicheng; QIN, Yong; SUN, Ruiqi; ZHOU, Xin; WANG, Enzhi; DONG, Xiaohang. **Better Zero-Shot Reasoning with Role-Play Prompting**. arXiv, 2024. Disponível em: <<http://arxiv.org/abs/2308.07702>>. Acesso em: 25/jun./25

KOTEK, Hadas; DOCKUM, Rikker; SUN, David Q. **Gender bias and stereotypes in Large Language Models**. In: *Proceedings of The ACM Collective Intelligence Conference*, 2023. [no place]: [no publisher]. 2023. p. 12–24. Disponível em: <<http://arxiv.org/abs/2308.14921>>. Acesso em: 01/set./25.

KUMAR, Yulia; MANIKANDAN, Anjana; LI, J. Jenny; MORREALE, Patricia. **Optimizing Large Language Models for Auto-Generation of Programming Quizzes**. In: *2024 IEEE Integrated STEM Education Conference (ISEC)*, 2024, Princeton, NJ, USA. Princeton, NJ, USA: IEEE. 2024. p. 1–5. Disponível em: <<https://ieeexplore.ieee.org/document/10665141/>>. Acesso em: 04/fev./25.

LEE, Jinsook; HICKE, Yann; YU, Renzhe; BROOKS, Christopher; KIZILCEC, René F. **The life cycle of large language models in education: A framework for understanding sources of bias**. In: *British Journal of Educational Technology*, v. 55, nº 5, p. 1982–2002, 2024. ISSN: 1467-8535, DOI: 10.1111/bjet.13505.

LINN, Robert L. **Accountability: Responsibility and Reasonable Expectations**. In: *Educational Researcher*, v. 32, nº 7, p. 3–13, 2003. ISSN: 0013-189X, DOI: 10.3102/0013189X032007003.

MAITY, Subhankar; DERROY, Aniket; SARKAR, Sudeshna. **A Novel Multi-Stage Prompting Approach for Language Agnostic MCQ Generation Using GPT**. In: *Advances in Information Retrieval*, 2024a, Cham. Cham: Springer Nature Switzerland.

2024. p. 268–277.

Subhankar Maity, Aniket Deroy, Sudeshna Sarkar. **Can large language models meet the challenge of generating school-level questions?** *In: Computers and Education: Artificial Intelligence*, v. 8, p. 100370, 2025. ISSN: 2666-920X, DOI: 10.1016/j.caeai.2025.100370.

MORAN, José Manuel. **A educação que desejamos: novos desafios e como chegar lá.** [no place]: Papirus Editora, 2007. 184 p. ISBN: 978-85-308-0835-8.

MULLA, Nikahat; GHARPURE, Prachi. **Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications.** *In: Progress in Artificial Intelligence*, v. 12, nº 1, p. 1–32, 2023. ISSN: 2192-6360, DOI: 10.1007/s13748-023-00295-9.

NAGY, Philip. **The Three Roles of Assessment: Gatekeeping, Accountability, and Instructional Diagnosis.** *In: Canadian Journal of Education / Revue canadienne de l'éducation*, v. 25, nº 4, p. 262–279, 2000. ISSN: 0380-2361, DOI: 10.2307/1585850.

NENKOVA, Ani; PASSONNEAU, Rebecca. **Evaluating Content Selection in Summarization: The Pyramid Method.** *In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, 2004, Boston, Massachusetts, USA. Boston, Massachusetts, USA: Association for Computational Linguistics. 2004. p. 145–152. Disponível em: <<https://aclanthology.org/N04-1019/>>. Acesso em: 19/maio/25.

OLIVEIRA, Laize Almeida De; SANTOS, Antonio Marques Dos; MARTINS, Rafael Castelo Guedes; OLIVEIRA, Erlania Lima De. **Inteligência artificial na educação: uma revisão integrativa da literatura.** *In: Peer Review*, v. 5, nº 24, p. 248–268, 2023. ISSN: 1541-1389, DOI: 10.53660/1369.prw2905.

PANE, John F.; STEINER, Elizabeth D.; BAIRD, Matthew D.; HAMILTON, Laura S. **Continued Progress: Promising Evidence on Personalized Learning.** [no place]: RAND Corporation, 2015. Acesso em: 22/abr./25.

PAPERT, Seymour. **Mindstorms—Children, Computers and Powerful Ideas.** [no place]: Basic Books, 1993. 252 p. ISBN: 0-465-04674-6. Disponível em: <<https://www.scirp.org/reference/referencespapers?referenceid=1468786>>. Acesso em: 03/jul./25.

PICÃO, Fábio Fornazieri; GOMES, Lucas Ferreira; ALVES, Luciene; BARPI, Odinei; LUCCHETTI, Tatiane Alves. **INTELIGÊNCIA ARTIFICIAL E EDUCAÇÃO: COMO A IA ESTÁ MUDANDO A MANEIRA COMO APRENDEMOS E ENSINAMOS**. In: *Revista Amor Mundi*, v. 4, nº 5, p. 197–201, 2023. ISSN: 2675-9152, DOI: 10.46550/amormundi.v4i5.254.

REYNOLDS, Laria; MCDONELL, Kyle. **Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm**. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI EA '21, 2021, New York, NY, USA. New York, NY, USA: Association for Computing Machinery. 2021. p. 1–7. Disponível em: <<https://doi.org/10.1145/3411763.3451760>>. Acesso em: 22/abr./25.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial: Uma Abordagem Moderna, 4ª edição Global**. 2004. Disponível em: <<https://aima.cs.berkeley.edu/global-index.html>>. Acesso em: 15/abr./24

SANTOS, Ramiro de Vasconcelos Dos; COELHO, João Vitor Venceslau; CACHO, Nelio Alessandro Azevedo; ARAÚJO, Daniel Sabino Amorim DE. **A criminal macrocause classification model: An enhancement for violent crime analysis considering an unbalanced dataset**. In: *Expert Systems with Applications*, v. 238, p. 121702, 2024. ISSN: 0957-4174, DOI: 10.1016/j.eswa.2023.121702.

SASANI, Ava. **As AI tools get smarter, they're growing more covertly racist, experts find**. In: *The Guardian*. 2024. Disponível em: <<https://www.theguardian.com/technology/2024/mar/16/ai-racism-chatgpt-gemini-bias>>. Acesso em: 01/set./25

SCRIVEN, Michael. **Evaluation thesaurus**. [no place]: Sage, 1991. Disponível em: <https://books.google.com/books?hl=pt-BR&lr=&id=koL0Fs_ZSvQC&oi=fnd&pg=PR7&dq=Michael+Scriven,+1991+%E2%80%93+Evaluation+Thesaurus&ots=Kb_s6XJghz&sig=eBeceC8_diSzq65SUEQF3r_4Vds>. Acesso em: 25/jun./25.

SIEMENS, George. **Connectivism: A Learning Theory for the Digital Age**. In: *International Journal of Instructional Technology and Distance Learning*, 2004. Disponível em: <<https://www.semanticscholar.org/paper/Connectivism%3A-A-Learning-Theory-for-the-Digital-Age-Siemens/7c7dd6c900c031b3685c761c72ebafdf3004caed>>. Acesso em: 03/set./25.

SILVA COSTA, Leandro; LOPES SOARES SILVA, Lenina; CASTRO SANTOS, Pablo;

KARINNE DE MOURA SARAIVA, Ana. **EDUCAÇÃO E INTELIGÊNCIA ARTIFICIAL: REFLEXÕES CRÍTICAS E PROPOSITIVAS**. In: *Revista Interinstitucional Artes de Educar*, v. 11, nº 1, p. 329–354, 2025. ISSN: 2359-6856, DOI: 10.12957/riae.2024.85857.

SOCIEDADE, Educação &. **Vieses algorítmicos da Inteligência Artificial e seus impactos na educação | SciELO em Perspectiva: Humanas**. 2025. Disponível em: <<https://humanas.blog.scielo.org/blog/2025/05/08/vieses-algoritmicos-da-inteligencia-artificial-e-seus-impactos-na-educacao/>>. Acesso em: 01/set./25

STEUER, Tim; BONGARD, Leonard; UHLIG, Jan; ZIMMER, Gianluca. **On the Linguistic and Pedagogical Quality of Automatic Question Generation via Neural Machine Translation**. In: *Technology-Enhanced Learning for a Free, Safe, and Sustainable World*, 2021, Cham. Cham: Springer International Publishing. 2021. p. 289–294.

Subhankar Maity, Aniket Deroy, and Sudeshna Sarkar. **Harnessing the Power of Prompt-based Techniques for Generating School-Level Questions using Large Language Models**. In: *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation. FIRE '23*, 2024b, New York, NY, USA. New York, NY, USA: Association for Computing Machinery. 2024. p. 30–39. Disponível em: <<https://dl.acm.org/doi/10.1145/3632754.3632755>>. Acesso em: 19/maio/25.

SWELLER, John. **Cognitive load theory, learning difficulty, and instructional design**. In: *Learning and Instruction*, v. 4, nº 4, p. 295–312, 1994. ISSN: 0959-4752, DOI: 10.1016/0959-4752(94)90003-5.

THELEN, Herbert A.; TYLER, Ralph W. **Implications for improving instruction in the high school**. 1950. Disponível em: <<https://psycnet.apa.org/record/2009-01944-013>>. Acesso em: 25/jun./25.

TRAN, Andrew; ANGELIKAS, Kenneth; RAMA, Egi; OKECHUKWU, Chiku; SMITH, David H.; MACNEIL, Stephen. **Generating Multiple Choice Questions for Computing Courses Using Large Language Models**. In: *2023 IEEE Frontiers in Education Conference (FIE)*, 2023, College Station, TX, USA. College Station, TX, USA: IEEE. 2023. p. 1–8. Disponível em: <<https://ieeexplore.ieee.org/document/10342898/>>. Acesso em: 04/fev./25.

VARGAS, Cláudio Pellini. **TARDIF, M. Saberes docentes e formação profissional. Petrópolis, R.J.: Editora Vozes, 2002**. In: *Instrumento: Revista de Estudo e Pesquisa*

em *Educação*, v. 10, 2008. ISSN: 1984-5499. Disponível em: <<https://periodicos.ufjf.br/index.php/revistainstrumento/article/view/18638>>. Acesso em: 08/jul./24.

WANG, Xu; FAN, Simin; HOUGHTON, Jessica; WANG, Lu. **Towards Process-Oriented, Modular, and Versatile Question Generation that Meets Educational Needs**. arXiv, 2022. Disponível em: <<http://arxiv.org/abs/2205.00355>>. Acesso em: 29/abr./25

WEI, Jason; WANG, Xuezhi; SCHUURMANS, Dale; BOSMA, Maarten; ICHTER, Brian; XIA, Fei; CHI, Ed; LE, Quoc; ZHOU, Denny. **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**. arXiv, 2023. Disponível em: <<http://arxiv.org/abs/2201.11903>>. Acesso em: 25/jun./25

WEISSBURG, Iain; ANAND, Sathvika; LEVY, Sharon; JEONG, Haewon. **LLMs are Biased Teachers: Evaluating LLM Bias in Personalized Education**. arXiv, 2025. Disponível em: <<http://arxiv.org/abs/2410.14012>>. Acesso em: 01/set./25

YANG, Jingfeng; JIN, Hongye; TANG, Ruixiang; HAN, Xiaotian; FENG, Qizhang; JIANG, Haoming; YIN, Bing; HU, Xia. **Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond**. arXiv, 2023. Disponível em: <<http://arxiv.org/abs/2304.13712>>. Acesso em: 28/jun./24

ZHANG, Ruqing; GUO, Jiafeng; CHEN, Lu; FAN, Yixing; CHENG, Xueqi. **A Review on Question Generation from Natural Language Text**. In: *ACM Trans. Inf. Syst.*, v. 40, n° 1, p. 14:1-14:43, 2021. ISSN: 1046-8188, DOI: 10.1145/3468889.

The Association for the Advancement of Artificial Intelligence. In: *AAAI*. [sem data]. Disponível em: <<https://aaai.org/>>. Acesso em: 28/jun./24

Apêndice A – Questões criadas e *Prompts* usados para os testes preliminares

Tabela A.1 - Testes preliminares: Questões Geradas com o *Prompt*.

LLMs	Questões criadas
Gemini	Olá! Claro, posso te ajudar com isso. Criarei cinco questões de Língua Portuguesa para alunos do sétimo ano, seguindo os níveis da Taxonomia de Bloom. Aqui estão elas: 1. (Nível: Conhecimento) Quais são as classes gramaticais que você conhece? Cite pelo menos cinco delas. 2. (Nível: Compreensão) Explique com suas palavras a diferença entre um substantivo próprio e um substantivo comum. Dê um exemplo de cada. 3. (Nível: Aplicação) Imagine que você está escrevendo um conto sobre um animal fantástico. Utilize pelo menos três adjetivos diferentes para descrever esse animal e explique por que você escolheu cada um deles. 4. (Nível: Análise) Leia a frase: "A garota esperta rapidamente resolveu o problema difícil." Identifique os elementos que indicam o gênero e o número das palavras "garota", "esperta" e "difícil". O que você pode concluir sobre a concordância nominal nessa frase? 5. (Nível: Avaliação) Em sua opinião, qual a importância de se utilizar a norma padrão da Língua Portuguesa em situações formais de comunicação? Justifique sua resposta com argumentos. Espero que essas questões sejam úteis para seus alunos do sétimo ano! Se precisar de mais alguma coisa, é só me dizer.
DeepSeek	Aqui estão cinco questões sobre Língua Portuguesa para alunos do 7º ano, alinhadas com os níveis da Taxonomia de Bloom (do mais básico ao mais complexo): 1. Lembrar (Nível 1 – Conhecimento) Questão: Liste os três tipos de discurso presentes em um texto narrativo e explique brevemente cada um deles. 2. Compreender (Nível 2 – Compreensão) Questão: Reescreva a frase abaixo substituindo o vocábulo "disse" por sinônimos mais expressivos, de acordo com a ação descrita: "Ele disse que estava com medo do cachorro." (Use verbos como "gritou", "sussurrou", "confessou", etc.) 3. Aplicar (Nível 3 – Aplicação)

	Questão: Identifique a figura de linguagem presente na frase abaixo e justifique sua resposta: <i>"O tempo é um rio que não para de fluir."</i> 4. Analisar (Nível 4 – Análise) Questão: Compare os textos <i>"Notícia de jornal"</i> e <i>"Conto literário"</i>, destacando duas diferenças estruturais e duas diferenças de linguagem entre eles. 5. Avaliar (Nível 5 – Avaliação) Questão: Leia o seguinte trecho de uma crônica e responda: O autor consegue transmitir uma crítica social de forma eficaz? Justifique sua resposta com elementos do texto. <i>"No ônibus lotado, ninguém ofereceu lugar à senhora idosa. Todos estavam ocupados demais com seus celulares, como se a vida real não existisse."</i> Essas questões abrangem diferentes habilidades cognitivas, desde a memorização até a análise crítica, seguindo a Taxonomia de Bloom. Você pode adaptar os textos ou temas conforme o conteúdo trabalhado em sala.
Claude	Questões de Língua Portuguesa - 7º Ano (Taxonomia de Bloom) Questão 1 - Nível Conhecimento (Recordar/Lembrar) Leia a tirinha abaixo: ![Tirinha de quadrinhos] Identifique o tipo de pronome destacado na frase "ESTE caderno está todo rasgado". a) Pronome possessivo b) Pronome demonstrativo c) Pronome indefinido d) Pronome pessoal Questão 2 - Nível Compreensão (Entender) Leia o trecho de texto abaixo: "A garota caminhava lentamente pela praia deserta. O sol já se punha no horizonte, e as ondas quebravam suavemente na areia. Ela pensava em tudo o que havia acontecido naquele dia." Explique com suas palavras qual é o tempo verbal predominante no texto e qual o efeito que esse tempo verbal produz na narrativa.

	Questão 3 - Nível Aplicação (Aplicar) Reescreva as frases abaixo, substituindo as palavras destacadas por sinônimos adequados, mantendo o mesmo sentido original: a) O aluno ficou FELIZ com a nota da prova. b) A CASA da minha avó fica no centro da cidade. c) O filme foi muito INTERESSANTE. Questão 4 - Nível Análise (Analisar) Leia o poema abaixo: "No meio do caminho tinha uma pedra Tinha uma pedra no meio do caminho Tinha uma pedra No meio do caminho tinha uma pedra" (Carlos Drummond de Andrade) Análise o recurso de linguagem utilizado pelo poeta neste trecho e explique qual é o efeito que ele produz no leitor. Em sua resposta, identifique esse recurso e justifique sua análise com elementos do texto. Questão 5 - Nível Avaliação/Criação (Avaliar/Criar) Observe a imagem de uma campanha publicitária contra o bullying nas escolas: ![[Imagem de campanha contra bullying]] Agora, crie um pequeno texto argumentativo (de 10 a 15 linhas) sobre a importância de combater o bullying nas escolas. Em seu texto, você deve: • Apresentar pelo menos dois argumentos convincentes • Utilizar linguagem adequada ao público adolescente • Propor ao menos uma solução para o problema • Concluir com uma reflexão que incentive a mudança de comportamento
Mistral	Claro! Aqui estão cinco questões sobre Língua Portuguesa para alunos do sétimo ano, seguindo os níveis da Taxonomia de Bloom: 1. Conhecimento (Lembrar): Quais são as classes gramaticais das palavras destacadas na seguinte frase: "O gato preto pulou rapidamente sobre o muro"? 2. Compreensão (Entender): Explique, com suas próprias palavras, o significado da seguinte metáfora: "Seus olhos eram estrelas brilhantes no céu da noite."

	3. Aplicação (Aplicar): Reescreva a seguinte frase no tempo futuro do pretérito: "Ela canta uma bela canção todas as manhãs." 4. Análise (Analisar): Compare e contraste os seguintes pares de palavras, destacando suas diferenças de significado e uso: "emigrante" e "imigrante". 5. Síntese (Criar): Crie um pequeno poema de quatro versos utilizando as seguintes palavras: "lua", "mar", "sonho" e "liberdade". Essas questões abrangem diferentes níveis cognitivos, desde a simples recordação de informações até a criação de novos conteúdos.
ChatGPT	Claro! Abaixo estão cinco questões de Língua Portuguesa para alunos do 7º ano, organizadas conforme os níveis da Taxonomia de Bloom (revisada por Anderson e Krathwohl):  1. Lembrar (Memorizar/Reconhecer) Pergunta: Qual é a função do sujeito em uma oração? Alternativas: A) Indicar o tempo verbal B) Identificar quem ou o que realiza a ação C) Mostrar a relação entre dois verbos D) Indicar o complemento do verbo Resposta correta: B  2. Compreender (Interpretar/Resumir) Pergunta: Leia o trecho abaixo: "A menina observava as estrelas com atenção." Qual é o significado da palavra "observava" no contexto da frase? Alternativas: A) Gritava com as estrelas B) Ignorava as estrelas C) Olhava com cuidado D) Corria em direção às estrelas Resposta correta: C  3. Aplicar (Usar regras ou conhecimento em novas situações) Pergunta: Transforme a frase "Os alunos estudam para a prova." na voz passiva. Resposta esperada: A prova é estudada pelos alunos.

	 4. Analisar (Comparar, Organizar, Distinguir partes) Pergunta: Leia as frases: "Ele chegou cedo." "Chegou cedo ele." Qual das frases está correta de acordo com a norma padrão da Língua Portuguesa? Explique por quê. Resposta esperada: A frase 1 está correta. Na norma padrão, a ordem direta é mais adequada (sujeito + verbo + complemento). Na frase 2, a inversão prejudica a clareza.  5. Criar (Planejar, Produzir, Inventar) Pergunta: Escreva um pequeno parágrafo (3 a 5 linhas) narrando uma situação usando pelo menos um verbo no pretérito perfeito e um adjetivo. Critérios de correção: • Uso correto do tempo verbal • Uso adequado de um adjetivo • Clareza e coesão textual
--	---

Tabela A.2 - Testes preliminares Meta – Prompt criados para gerar as questões.

LLMs	Questões criadas
Gemini	"Crie cinco questões de múltipla escolha sobre o conteúdo de [inserir um tópico específico de Língua Portuguesa do 7º ano aqui, por exemplo: verbos, substantivos, adjetivos, concordância verbal, etc.]. Cada questão deve corresponder a um nível diferente da Taxonomia de Bloom, conforme descrito abaixo: 1. Memorização: Elabore uma questão que exija a recordação de informações básicas e definições sobre o tópico. 2. Compreensão: Crie uma questão que demande a capacidade de entender o significado do conteúdo e expressá-lo de outra forma. 3. Aplicação: Desenvolva uma questão que envolva a utilização do conhecimento em uma situação nova ou concreta. 4. Análise: Formule uma questão que requeira a decomposição do conteúdo em partes menores e a identificação das relações entre elas. 5. Avaliação: Apresente uma questão que peça a emissão de um julgamento sobre o valor do conteúdo, baseando-se em critérios específicos. Para cada questão, forneça: • O enunciado da questão.

	• Quatro opções de resposta (A, B, C, D), sendo apenas uma correta. • A identificação do nível da Taxonomia de Bloom a que a questão pertence." Exemplo de como usar o prompt (substituindo o tópico): "Crie cinco questões de múltipla escolha sobre o conteúdo de verbos para alunos do sétimo ano. Cada questão deve corresponder a um nível diferente da Taxonomia de Bloom, conforme descrito abaixo: 1. Memorização: Elabore uma questão que exija a recordação de informações básicas e definições sobre o tópico. 2. Compreensão: Crie uma questão que demande a capacidade de entender o significado do conteúdo e expressá-lo de outra forma. 3. Aplicação: Desenvolva uma questão que envolva a utilização do conhecimento em uma situação nova ou concreta. 4. Análise: Formule uma questão que requeira a decomposição do conteúdo em partes menores e a identificação das relações entre elas. 5. Avaliação: Apresente uma questão que peça a emissão de um julgamento sobre o valor do conteúdo, baseando-se em critérios específicos. Para cada questão, forneça: • O enunciado da questão. • Quatro opções de resposta (A, B, C, D), sendo apenas uma correta. • A identificação do nível da Taxonomia de Bloom a que a questão pertence." Ao fornecer esse prompt, você estará direcionando a criação de questões que abrangem diferentes habilidades cognitivas, desde a simples memorização até a avaliação crítica, o que é fundamental para uma avaliação completa do aprendizado dos alunos. Lembre-se de especificar o tópico de Língua Portuguesa desejado para obter questões mais direcionadas.
DeepSeek	"Crie cinco questões sobre Língua Portuguesa para alunos do sétimo ano, abrangendo os seguintes níveis da Taxonomia de Bloom: 1. Lembrar (Nível 1): Questão que avalie a memorização de conceitos básicos (ex.: definições, regras gramaticais). 2. Compreender (Nível 2): Questão que exija a explicação ou interpretação de um conteúdo (ex.: significado de um texto, analogias). 3. Aplicar (Nível 3): Questão que peça para usar um conhecimento em uma situação prática (ex.: produção de frases com determinada regra gramatical). 4. Analisar (Nível 4): Questão que solicite a comparação,

	classificação ou identificação de padrões (ex.: diferenças entre figuras de linguagem). 5. Avaliar (Nível 5): Questão que envolva argumentação ou crítica com base em evidências (ex.: opinar sobre efeitos de recursos linguísticos em um texto). Instruções: • Foque em conteúdos do 7º ano (ex.: classes gramaticais, concordância, gêneros textuais, figuras de linguagem). • Inclua um mix de perguntas objetivas e dissertativas. • Contextualize as questões com exemplos ou situações do cotidiano quando possível. Exemplo de saída esperada: 1. (Lembrar) Liste três pronomes pessoais do caso reto. 2. (Compreender) Explique com suas palavras o que é uma metáfora. 3. (Aplicar) Reescreva a frase 'Ontem, eu vai ao parque' corrigindo o erro de concordância. 4. (Analisar) Compare o uso da linguagem formal e informal em uma notícia e em uma mensagem entre amigos. 5. (Avaliar) Justifique por que o uso de hipérbole pode ser eficaz em um anúncio publicitário." Adaptações possíveis: • Para incluir o nível Criar (Nível 6), adicione uma questão como: <i>"Crie um diálogo entre personagens usando duas figuras de linguagem estudadas."</i> • Substitua os tópicos gramaticais por outros conforme o currículo (ex.: variação linguística, funções da linguagem). Esse prompt garante questões diversificadas e alinhadas com habilidades cognitivas progressivas.
Claude	Prompt para Criação de Questões de Língua Portuguesa (7º ano) baseadas na Taxonomia de Bloom Orientações Gerais Crie cinco questões de Língua Portuguesa para alunos do 7º ano do Ensino Fundamental. Cada questão deve corresponder a um nível diferente da Taxonomia de Bloom, seguindo uma progressão do nível mais básico ao mais complexo. As questões devem ser adequadas à faixa etária (12-13 anos) e alinhadas com o currículo de Língua Portuguesa para essa série. Conteúdos Sugeridos • Classes gramaticais (substantivos, adjetivos, verbos, advérbios) • Interpretação de textos narrativos, descritivos e argumentativos • Figuras de linguagem básicas (metáfora, comparação,

	personificação) • Ortografia e acentuação • Coesão e coerência textual • Variação linguística Estrutura das Questões (por nível da Taxonomia de Bloom) Questão 1: Nível "Lembrar" (Conhecimento) Crie uma questão que exija que o aluno apenas reconheça, identifique ou recorde informações específicas sobre o conteúdo. Pode ser uma questão de múltipla escolha sobre identificação de classes gramaticais ou regras ortográficas. Questão 2: Nível "Entender" (Compreensão) Formule uma questão que peça ao aluno para demonstrar compreensão do conteúdo, explicando conceitos com suas próprias palavras ou interpretando textos simples. Pode envolver a compreensão do sentido de palavras em contexto ou identificação da ideia principal de um pequeno parágrafo. Questão 3: Nível "Aplicar" (Aplicação) Desenvolva uma questão onde o aluno precise aplicar conhecimentos adquiridos para resolver um problema novo. Por exemplo, reescrever frases corrigindo erros gramaticais ou completar textos com palavras adequadas ao contexto. Questão 4: Nível "Analisar" (Análise) Elabore uma questão que exija que o aluno decomponha o conteúdo em partes e entenda a relação entre elas. Pode ser uma análise de texto para identificar recursos estilísticos ou elementos de coesão textual. Questão 5: Nível "Avaliar" e "Criar" (Avaliação e Síntese) Construa uma questão que combine os níveis mais altos da taxonomia, pedindo que o aluno faça julgamentos baseados em critérios específicos e/ou crie algo novo a partir do conhecimento adquirido. Por exemplo, uma produção textual curta ou reescrita criativa de um trecho com propósito definido. Formato de Entrega • Forneça cada questão claramente numerada • Inclua a identificação do nível da Taxonomia de Bloom para cada questão
--	---

	• Anexe as respostas ou critérios de correção para cada questão • Adicione uma breve justificativa pedagógica para cada questão, explicando como ela atende ao nível cognitivo pretendido
Mistral	Crie cinco questões sobre Língua Portuguesa para alunos do sétimo ano, seguindo os níveis da Taxonomia de Bloom. Cada questão deve ser elaborada para avaliar um nível diferente da taxonomia, conforme descrito abaixo: 1. Conhecimento (Lembrar): Crie uma questão que avalie a capacidade do aluno de lembrar informações básicas sobre um tema específico de Língua Portuguesa, como classes de palavras ou regras gramaticais. 2. Compreensão (Entender): Elabore uma questão que exija que o aluno demonstre entendimento de um conceito, como a interpretação de um texto ou a explicação de uma figura de linguagem. 3. Aplicação (Aplicar): Desenvolva uma questão que peça ao aluno para aplicar um conceito aprendido em uma nova situação, como a criação de uma frase usando uma determinada figura de linguagem. 4. Análise (Analisar): Crie uma questão que solicite ao aluno que analise um texto ou uma estrutura linguística, identificando partes, relações ou princípios de organização. 5. Síntese (Criar): Elabore uma questão que incentive o aluno a criar algo novo a partir do conhecimento adquirido, como a escrita de um pequeno texto ou a criação de um diálogo usando determinadas regras gramaticais. Certifique-se de que as questões sejam adequadas ao nível de conhecimento e habilidade dos alunos do sétimo ano e que cubram uma variedade de tópicos dentro da disciplina de Língua Portuguesa.
ChatGPT	Prompt: Geração de Questões de Língua Portuguesa com Base na Taxonomia de Bloom Crie cinco questões de Língua Portuguesa destinadas a alunos do 7º ano do Ensino Fundamental, abordando conteúdos comuns dessa série (como interpretação de texto, ortografia, gramática, pontuação, figuras de linguagem etc.). Cada questão deve corresponder a um nível da Taxonomia de Bloom (revisada), sendo eles: 1. Lembrar (recordar fatos ou conceitos) 2. Compreender (interpretar, resumir ou explicar) 3. Aplicar (usar o conhecimento em novas situações) 4. Analisar (distinguir partes, inferir significados) 5. Avaliar ou Criar (julgar com base em critérios ou produzir algo novo)

	Para cada questão, forneça: • Número da questão • Habilidade cognitiva correspondente • Enunciado claro e adequado à faixa etária • Alternativas (quando for questão de múltipla escolha) ou estrutura de resposta (quando discursiva) • Gabarito e justificativa Formato de saída esperado: **Questão 1 — Lembrar** [Enunciado] a) ... b) ... c) ... d) ... **Gabarito:** letra X — [justificativa] **Questão 2 — Compreender** Certifique-se de manter uma linguagem compatível com alunos de 12 a 13 anos e temas que façam sentido no contexto escolar dessa faixa etária.
--	---

Tabela A.3 - Testes preliminares questões geradas com o Meta - Prompt.

LLMs	Questões criadas
Gemini	Com certeza! Vamos criar cinco questões de múltipla escolha sobre o conteúdo de substantivos para alunos do sétimo ano, cada uma correspondendo a um nível diferente da Taxonomia de Bloom. 1. Memorização: • Enunciado: Qual das alternativas abaixo apresenta a definição correta de substantivo? ○ (A) Palavra que indica uma ação, estado ou fenômeno da natureza. ○ (B) Palavra que acompanha o substantivo, modificando seu sentido. ○ (C) Palavra que nomeia seres, objetos, lugares, sentimentos, etc. ○ (D) Palavra que liga duas orações ou termos de mesma função.

- **Opção Correta:** (C)
- **Nível da Taxonomia de Bloom:** Memorização

2. Compreensão:

- **Enunciado:** Na frase "A alegria contagiante tomou conta da festa", qual palavra expressa um sentimento e funciona como substantivo?
 - (A) A
 - (B) Contagiant
 - (C) Tomou
 - (D) Alegria

Opção Correta: (D)

Nível da Taxonomia de Bloom: Compreensão

3. Aplicação:

Enunciado: Em qual das frases abaixo a palavra destacada é classificada como um substantivo próprio?

- O **cachorro** latiu durante a noite.
- A **cidade** estava movimentada.
- **Maria** foi ao mercado pela manhã.
- Comprei um livro **interessante**.

Opção Correta: (C)

Nível da Taxonomia de Bloom: Aplicação

4. Análise:

Enunciado: Leia a frase: "Os bravos guerreiros defenderam a aldeia com coragem." Identifique e classifique os substantivos presentes na frase, considerando se são comuns ou próprios, concretos ou abstratos. Qual das alternativas apresenta a análise correta?

- Guerreiros (comum, concreto), aldeia (comum, concreto), coragem (comum, abstrato).
- Os (artigo), bravos (adjetivo), guerreiros (próprio, concreto), defenderam (verbo), a (artigo), aldeia (comum, concreto), com (preposição), coragem (comum, abstrato).
- Guerreiros (comum, abstrato), aldeia (comum, concreto), coragem (comum, concreto).
- Guerreiros (próprio, abstrato), aldeia (próprio, concreto), coragem

	(comum, abstrato). Opção Correta: (A) Nível da Taxonomia de Bloom: Análise 5. Avaliação: Enunciado: Considere as seguintes palavras: <i>tempo, amizade, mesa, Brasil</i>. Qual critério você utilizaria para agrupar essas palavras em categorias de substantivos concretos e abstratos? Avalie a validade desse critério e escolha a alternativa que melhor justifica a sua escolha. ▪ O critério é inválido, pois todos os substantivos nomeiam coisas que podemos ver e tocar. ▪ O critério é válido, pois permite distinguir entre substantivos que dependem de um ser para existir (<i>tempo, amizade</i>) e aqueles que existem por si só (<i>mesa, Brasil</i>). ▪ O critério é irrelevante, pois a classificação de substantivos não interfere na compreensão da língua portuguesa. ▪ O critério é parcialmente válido, mas a palavra "Brasil" deveria ser classificada como abstrato por se referir a uma nação. Opção Correta: (B) Nível da Taxonomia de Bloom: Avaliação
DeepSeek	1. Lembrar (Nível 1) Objetiva Marque a alternativa que lista apenas conjunções coordenativas: a) Mas, e, porque, logo. b) Porém, ou, pois, portanto. c) E, mas, ou, nem. d) Quando, que, se, como. (Conteúdo: Classes gramaticais - conjunções) 2. Compreender (Nível 2) Dissertativa Leia o trecho: "O tempo é um rio que não volta atrás." Explique, com suas palavras, qual figura de linguagem está presente e qual é seu significado. (Conteúdo: Figuras de linguagem - metáfora) 3. Aplicar (Nível 3) Prática Reescreva a frase abaixo corrigindo o erro de concordância verbal: "As menina foi ao cinema ontem." (Conteúdo: Concordância verbal)

	4. Analisar (Nível 4) Dissertativa Compare os dois textos abaixo, indicando qual é formal e qual é informal. Justifique sua resposta com características de cada um: Texto A: "Caro Sr. Silva, solicitamos o envio dos documentos até sexta-feira." Texto B: "Oi, cara! Manda os docs até sexta, tá bom?" (Conteúdo: Variação linguística - registros formal e informal) 5. Avaliar (Nível 5) Argumentativa Um anúncio de loja diz: "Os preços mais baixos do universo!". Você acha que o uso da hipérbole é eficaz nesse caso? Justifique sua opinião com base no efeito persuasivo desse recurso. (Conteúdo: Figuras de linguagem - hipérbole e funções da linguagem) Bônus: Criar (Nível 6) Produção textual Crie um diálogo curto (4-5 linhas) entre dois amigos usando pelo menos duas figuras de linguagem estudadas (ex.: ironia, personificação, comparação). (Conteúdo: Figuras de linguagem e produção textual)
Claude	Questões de Língua Portuguesa para 7º ano baseadas na Taxonomia de Bloom Vou criar cinco questões progressivas seguindo os diferentes níveis da Taxonomia de Bloom, adequadas para alunos do 7º ano do Ensino Fundamental. Questão 1: Nível "Lembrar" (Conhecimento) Classifique as palavras destacadas na frase abaixo quanto à sua classe gramatical: "A menina correu alegremente pelo parque ensolarado." a) substantivo - verbo - advérbio - substantivo - adjetivo b) adjetivo - verbo - advérbio - substantivo - adjetivo c) substantivo - substantivo - advérbio - substantivo - adjetivo d) substantivo - verbo - adjetivo - substantivo - adjetivo

Resposta: Alternativa A

Justificativa pedagógica: Esta questão trabalha no nível mais básico da taxonomia, exigindo que o aluno apenas reconheça e identifique classes gramaticais já estudadas. O exercício não demanda interpretação complexa, apenas a lembrança de conceitos gramaticais e a capacidade de identificá-los no contexto da frase.

Questão 2: Nível "Entender" (Compreensão)

Leia o texto abaixo e responda:

"Quando a chuva finalmente parou, as nuvens se afastaram como cortinas que se abrem no teatro. O sol, tímido no início, logo tomou conta do céu e começou a secar as poças d'água que se formaram nas calçadas."

Que figura de linguagem está presente no trecho "as nuvens se afastaram como cortinas que se abrem no teatro"? Explique com suas próprias palavras o que o autor quis dizer com essa expressão.

Resposta esperada: No trecho está presente a figura de linguagem chamada comparação, pois o autor compara o movimento das nuvens com cortinas que se abrem no teatro, utilizando a palavra "como" para estabelecer essa relação. O autor quis dizer que as nuvens foram se afastando gradualmente, abrindo espaço no céu de maneira semelhante às cortinas que se abrem em um teatro, revelando o palco (que neste caso seria o céu limpo).

Justificativa pedagógica: Esta questão avança para o nível de compreensão, pois o aluno precisa não apenas identificar a figura de linguagem (comparação), mas também compreender seu significado no contexto e explicar com suas próprias palavras, demonstrando entendimento do recurso estilístico utilizado pelo autor.

Questão 3: Nível "Aplicar" (Aplicação)

Reescreva as frases abaixo, substituindo as palavras repetidas por pronomes adequados e fazendo outras alterações necessárias para manter a coerência do texto:

"João comprou um livro novo. João começou a ler o livro assim que João chegou em casa. O livro era sobre aventuras no espaço. João mostrou o livro para seus amigos no dia seguinte."

Resposta esperada: João comprou um livro novo. Ele começou a lê-lo assim que chegou em casa. O livro era sobre aventuras no espaço. Ele o mostrou para seus amigos no dia seguinte.

Justificativa pedagógica:

Esta questão trabalha no nível de aplicação, pois exige que o aluno aplique seu conhecimento sobre pronomes e coesão textual em um contexto prático. O estudante precisa identificar as repetições, selecionar os pronomes adequados e fazer as modificações necessárias para manter o sentido e a fluência do texto.

Questão 4: Nível "Analisar" (Análise)**Leia o texto abaixo e analise os elementos de coesão:**

"Os adolescentes de hoje passam muito tempo nas redes sociais. Entretanto, nem sempre sabem usar essas ferramentas com responsabilidade. Muitos compartilham informações pessoais sem pensar nas consequências. Além disso, podem se expor a conteúdos inadequados. Por isso, é fundamental que os pais conversem com seus filhos sobre o uso consciente da internet."

- a) Identifique três elementos de coesão presentes no texto (palavras ou expressões que conectam as ideias).**
- b) Explique a função de cada um desses elementos, mostrando como eles contribuem para estabelecer as relações entre as frases.**
- c) Como o texto seria afetado se esses elementos fossem removidos?**

Resposta esperada:

a) Elementos de coesão: "Entretanto", "Além disso", "Por isso"

b) "Entretanto": Estabelece uma relação de oposição ou contraste entre a primeira e a segunda frase, indicando que há um problema no uso das redes sociais pelos adolescentes.

"Além disso": Adiciona uma nova informação ao texto, ampliando as consequências negativas do uso das redes sociais.

"Por isso": Introduce uma conclusão ou consequência lógica a partir das informações anteriores, apresentando uma solução para o problema.

c) Se esses elementos fossem removidos, as ideias ficariam desconectadas e o texto perderia sua progressão lógica. O leitor teria dificuldade para compreender as relações entre as frases, como a oposição, adição de informações e conclusão. O texto pareceria uma lista

de frases soltas, sem coerência entre elas.

Justificativa pedagógica:

Esta questão trabalha no nível de análise, pois o aluno precisa decompor o texto e examinar como os elementos de coesão funcionam para conectar as ideias. Além de identificar os elementos, o estudante deve analisar sua função e refletir sobre como eles contribuem para a construção do significado global do texto.

Questão 5: Nível "Avaliar" e "Criar" (Avaliação e Síntese)

Leia a manchete de jornal abaixo e realize as atividades propostas:

"Jovens abandonam livros e preferem redes sociais, aponta pesquisa."

a) Avalie criticamente essa manchete. Você concorda com essa afirmação? Justifique seu posicionamento com pelo menos dois argumentos.

b) Crie um pequeno texto argumentativo (entre 10-15 linhas) posicionando-se sobre esse tema. Seu texto deve incluir:

- Um título criativo
- Pelo menos dois argumentos que sustentem sua opinião
- Um contra-argumento
- Uma conclusão que proponha uma solução ou reflexão
- Uso adequado de conectivos para garantir a coesão textual

Critérios de avaliação:

a) Avaliação crítica:

- Posicionamento claro sobre a afirmação
- Argumentos coerentes e bem fundamentados
- Capacidade de análise crítica da manchete

b) Texto argumentativo:

- Título criativo e pertinente ao tema
- Estrutura adequada (introdução, desenvolvimento, conclusão)
- Argumentos bem desenvolvidos e articulados
- Presença de contra-argumento
- Proposta de solução ou reflexão final
- Uso adequado de conectivos e elementos de coesão
- Adequação à norma culta da língua portuguesa

	Justificativa pedagógica: Esta questão combina os níveis mais complexos da Taxonomia de Bloom (avaliação e criação). No item A, o aluno deve avaliar criticamente uma afirmação, fazendo julgamentos baseados em critérios e apresentando argumentos. No item B, o estudante precisa criar um texto próprio, sintetizando conhecimentos e aplicando habilidades de argumentação, coesão textual e norma culta. A atividade estimula o pensamento crítico e a capacidade de expressão escrita do aluno.
Mistral	Aqui estão cinco questões sobre Língua Portuguesa para alunos do sétimo ano, seguindo os níveis da Taxonomia de Bloom: 1. Conhecimento (Lembrar): Questão: Liste cinco classes de palavras (ou classes gramaticais) e dê um exemplo de cada uma. 2. Compreensão (Entender): Questão: Leia o seguinte trecho e explique o significado da metáfora usada: "O tempo é um ladrão silencioso." 3. Aplicação (Aplicar): Questão: Crie uma frase usando a figura de linguagem chamada de personificação, onde você atribui características humanas a um objeto inanimado. 4. Análise (Analisar): Questão: Analise o seguinte parágrafo e identifique o sujeito e o predicado em cada uma das frases: "O sol brilhava no céu azul. As crianças corriam pelo parque. Os pássaros cantavam nas árvores." 5. Síntese (Criar): Questão: Escreva um pequeno diálogo entre dois personagens usando, pelo menos, três diferentes figuras de linguagem (como metáfora, comparação, hipérbole, etc.). Certifique-se de identificar quais figuras de linguagem você usou. Essas questões cobrem uma variedade de tópicos e habilidades dentro da disciplina de Língua Portuguesa e são adequadas para alunos do sétimo ano.
ChatGPT	Claro! Abaixo estão cinco questões de Língua Portuguesa para alunos do 7º ano, alinhadas com os níveis da Taxonomia de Bloom (revisada), com foco em conteúdos comuns da série: Questão 1 — Lembrar

Habilidade: Recordar fatos ou conceitos (Ortografia)

Qual das palavras abaixo está escrita corretamente de acordo com a norma ortográfica da Língua Portuguesa?

- a) Enxergar
- b) Enchergar
- c) Enxergarh
- d) Enchergar

Gabarito: letra a — A forma correta e aceita na norma-padrão é "enxergar". As demais opções apresentam grafias incorretas.

Questão 2 — Compreender

Habilidade: Interpretar (Interpretação de texto)

Leia o trecho abaixo:

"O menino olhava a pipa subindo no céu como se fosse um pássaro colorido ganhando liberdade."

Qual o melhor significado da comparação feita no trecho acima?

- a) A pipa era feita com penas de pássaro.
- b) A pipa estava subindo lentamente.
- c) A pipa parecia um pássaro livre no céu.
- d) O menino queria ser um pássaro.

Gabarito: letra c — A comparação sugere que a pipa, ao subir, lembra um pássaro colorido ganhando liberdade, indicando leveza e liberdade no movimento.

Questão 3 — Aplicar

Habilidade: Usar o conhecimento em novas situações (Pontuação)

Reescreva a frase abaixo, inserindo corretamente as vírgulas onde forem necessárias:

"João Maria Pedro e Ana foram ao parque depois da escola"

Resposta esperada:

João, Maria, Pedro e Ana foram ao parque depois da escola.

Gabarito e justificativa: A vírgula é usada para separar os elementos de uma enumeração, tornando a frase mais clara. Sem as vírgulas, os nomes parecem um nome composto.

Questão 4 — Analisar

Habilidade: Distinguir partes, inferir significados (Figuras de linguagem)

	Leia a frase: "A cidade acordou triste naquele dia nublado." Qual figura de linguagem foi usada e por quê? a) Hipérbole — porque exagera o sentimento da cidade. b) Metáfora — porque compara a cidade a uma pessoa triste. c) Personificação — porque dá características humanas à cidade. d) Ironia — porque diz o contrário do que se pensa. Gabarito: letra c — A cidade não pode "acordar" nem "ficar triste", portanto a frase usa a personificação para atribuir sentimentos humanos à cidade. Questão 5 — Avaliar ou Criar Habilidade: Produzir algo novo (Criação de texto) Proposta: Crie uma frase usando a figura de linguagem chamada "metáfora" para descrever uma situação do seu dia a dia. Resposta esperada: Resposta aberta. Exemplos aceitáveis: • "Meu quarto é um campo de batalha depois do fim de semana." • "Minha cabeça era um vulcão prestes a explodir antes da prova." Critérios de correção: • Uso correto da metáfora (comparação implícita, sem uso de "como"); • Clareza e criatividade na construção da frase. Justificativa: Avalia a habilidade do aluno em compreender e aplicar o conceito de metáfora, criando algo original e coerente.
--	---

Segundo teste para realizar alterações no *Meta-Prompting*

Tabela A.4 - Meta – Prompt para a criação das questões.

LLMs	Questões criadas
Gemini	Com certeza! Como especialista em educação com profundo conhecimento da BNCC de Língua Portuguesa e da Taxonomia de Bloom, criarei um prompt detalhado e eficaz para gerar questões educacionais de alta qualidade para alunos do 7º ano do Ensino Fundamental.

Prompt Especializado para Geração de Questões de Língua Portuguesa (7º Ano EF)

Público-alvo: Alunos do 7º ano do Ensino Fundamental.

Objetivo: Gerar questões de Língua Portuguesa que sejam desafiadoras, engajadoras, contextualmente relevantes e alinhadas com as competências e habilidades da BNCC para o 7º ano, utilizando a Taxonomia de Bloom para garantir a progressão dos níveis cognitivos.

Estrutura do Prompt:

Você é um gerador de questões educacionais para o 7º ano do Ensino Fundamental. Sua tarefa é criar **X** questões de Língua Portuguesa, cobrindo os diferentes níveis da Taxonomia de Bloom (de *Lembrar a Criar*), e que estejam alinhadas às competências e habilidades específicas da BNCC para o 7º ano.

Diretrizes Essenciais:

1. Contextualização Relevante:

- As questões devem apresentar contextos significativos para adolescentes de 11 a 13 anos. Pense em temas como: redes sociais, memes, cultura pop (séries, filmes, músicas, jogos), bullying, questões ambientais, diversidade, notícias do cotidiano, esportes, amizade, família, escola, etc.
- Utilize situações-problema ou cenários que simulem situações reais de uso da língua.
- **Evitar:** Contextos muito infantis, desinteressantes ou excessivamente acadêmicos sem ligação com o universo do aluno.

2. Clareza e Objetividade nos Enunciados:

- Os enunciados devem ser concisos, diretos e sem ambiguidades.
- Apresentar uma única solicitação por questão, a menos que seja uma questão de múltipla escolha com subitens claros.
- Utilizar linguagem adequada à faixa etária, mas sem simplificar demais o vocabulário técnico da disciplina (ex: "substantivo", "verbo", "argumento", "tese").
-

3. Alinhamento com a BNCC (7º Ano - Língua Portuguesa):

- **Foco nas Habilidades e Objetos de Conhecimento:** Priorize as habilidades listadas para o 7º ano em eixos como:
 - **Leitura:** Análise de diferentes gêneros textuais

	(notícia, reportagem, meme, texto de opinião, tira, HQ, conto, poema, letra de música, propaganda), inferência de sentido, identificação de tese/argumentos, relação entre texto e imagem. ▪ Produção de Textos: Planejamento, textualização e revisão de textos de diferentes gêneros (cartas de leitor, artigos de opinião, resenhas, posts para redes sociais, contos). ▪ Oralidade: Participação em debates, apresentação de ideias, discussão de temas relevantes. ▪ Análise Linguística/Semiótica: Uso de elementos linguísticos (coesão, coerência), classes de palavras (substantivo, adjetivo, verbo, advérbio, preposição, conjunção), concordância verbal e nominal, regência, figuras de linguagem (metáfora, comparação, personificação, hipérbole), efeitos de sentido. ○ Competências Gerais e Específicas: As questões devem, implicitamente, mobilizar as competências gerais da BNCC (ex: pensamento crítico, comunicação, cultura digital, autoconhecimento e autocuidado) e as competências específicas de Língua Portuguesa (ex: compreender, analisar, produzir, utilizar diferentes linguagens). 4. Aplicação da Taxonomia de Bloom (Níveis Cognitivos): ○ Lembrar: Recordar fatos, termos, conceitos básicos. (Ex: "Qual a função de um substantivo no período?") ○ Compreender: Explicar ideias ou conceitos, interpretar informações. (Ex: "Com suas palavras, explique o que o autor quis dizer no segundo parágrafo.") ○ Aplicar: Usar informações em novas situações, resolver problemas. (Ex: "Identifique no texto um exemplo de advérbio de tempo e explique seu efeito de sentido.") ○ Analisar: Quebrar informações em partes, identificar motivos ou causas, fazer inferências. (Ex: "Compare o ponto de vista do autor do Texto A com o ponto de vista do autor do Texto B sobre o uso de redes sociais, apontando semelhanças e diferenças.") ○ Avaliar: Julgar o valor das informações, ideias ou argumentos, fazer escolhas justificadas. (Ex: "Você concorda com a afirmação principal do texto sobre o impacto da tecnologia na leitura? Justifique sua resposta com base em argumentos consistentes.") ○ Criar: Produzir um novo trabalho ou ideia original, combinar elementos de forma nova. (Ex: "A partir do cenário apresentado, escreva um pequeno trecho de um conto onde o personagem principal utiliza a metáfora para
--	---

descrever seus sentimentos.")

Formato das Questões (Variedade):

- **Múltipla Escolha:** Com 4 ou 5 alternativas, sendo uma correta e as demais bons distratores.
- **Dissertativas Curtas:** Que exigem respostas concisas e diretas.
- **Dissertativas Abertas:** Que permitem maior elaboração e desenvolvimento de argumentos.
- **Completar Lacunas:** Para testar conhecimentos específicos.
- **Associação de Colunas:** Para relacionar conceitos ou exemplos.
- **Questões com Imagens/Gráficos/Tabelas:** Que exijam a interpretação de elementos visuais.
- **Questões com Trechos de Áudio/Vídeo (se aplicável, descrever o conteúdo):** Para trabalhar oralidade ou a multimodalidade.
-

Formato da Resposta Esperada do Gerador de Questões:

Para cada questão gerada, forneça:

- **Número da Questão:**
- **Nível da Taxonomia de Bloom:** (Lembrar, Compreender, Aplicar, Analisar, Avaliar, Criar)
- **Habilidade(s) da BNCC (Código e Descrição Resumida):** (Ex: EF67LP04 - Análise de gêneros)
- **Contexto/Texto Base (se houver):** (Ex: Trecho de uma notícia, imagem de um meme, letra de música, etc. Se for um texto longo, indicar que o texto X seria fornecido)
- **Enunciado da Questão:**
- **Tipo de Questão:** (Múltipla Escolha, Dissertativa Curta, etc.)
- **Gabarito/Sugestão de Resposta (para Dissertativas, pontos-chave):** (Opcional, mas útil para o avaliador)
- **Justificativa (para Múltipla Escolha - breve explicação da alternativa correta e por que as outras são incorretas):** (Opcional)

Exemplo de Como o Gerador Deve Utilizar o Prompt:

"Gere 3 questões de Língua Portuguesa para o 7º ano do Ensino Fundamental. Uma questão deve ser do nível 'Compreender', focando em leitura e interpretação de gêneros textuais curtos (como um post de rede social). Outra, do nível 'Aplicar', envolvendo a identificação de classes de palavras em um contexto sobre bullying. E a última, do nível 'Avaliar', pedindo a opinião sobre um tema atual, com justificativa."

Este prompt abrangente e detalhado permitirá a criação de questões de alta qualidade, que realmente avaliem as competências e habilidades dos alunos do 7º ano de forma relevante e desafiadora.

Prompt para Geração de Questões Educacionais de Língua Portuguesa (7º Ano - BNCC e Taxonomia de Bloom)**Instruções:**

Você deve criar questões alinhadas à **BNCC de Língua Portuguesa (7º ano)** e à **Taxonomia de Bloom**, assegurando contextos significativos, clareza e relevância para estudantes de 12-13 anos.

Diretrizes para a Criação das Questões:**1. Alinhamento à BNCC:**

- Focar em habilidades como **EF69LPxx** (leitura/escrita/oralidade) e **EF07LPyy** (gramática/linguística), conforme os eixos:
 - **Leitura e compreensão textual** (gêneros como contos, crônicas, notícias, memes, infográficos).
 - **Produção textual** (narrativas, argumentações, textos multimodais).
 - **Análise linguística/semiótica** (variação linguística, figuras de linguagem, coerência/coesão).

2. Níveis Cognitivos (Taxonomia de Bloom Revisada):

- Incluir questões que explorem:
 - **Lembrar** (ex.: identificar elementos textuais).
 - **Compreender** (ex.: parafrasear, inferir).
 - **Aplicar** (ex.: usar conceitos em novas situações).
 - **Analisar** (ex.: comparar textos, estruturas).
 - **Avaliar** (ex.: julgar argumentos, revisar textos).
 - **Criar** (ex.: produzir um texto com base em diretrizes).

3. Contextualização Significativa:

- Usar temas próximos à realidade do aluno (redes sociais, desafios da adolescência, sustentabilidade, cultura pop).
- Incluir **textos autênticos** (trechos de livros juvenis, posts, charges) ou situações-problema.

4. Formatos Possíveis:

- Múltipla escolha com justificativa.
- Resposta construída (curta ou dissertativa).
- Análise de texto com comandos específicos (ex.: "Destaque metáforas no poema abaixo").
- Produção guiada (ex.: "Escreva um e-mail de reclamação sobre um produto defeituoso").

Exemplo de Questão:

Habilidade BNCC: EF07LP16 (Analisar efeitos de sentido de figuras de linguagem).

Nível Cognitivo: Analisar.

Contexto: Charge sobre uso excessivo de celulares.

"Na charge abaixo, o balão diz: 'Meu celular é meu melhor amigo!'. Que

	<i>figura de linguagem é utilizada e qual seu efeito de sentido no contexto?"</i> Solicitação Final: Gere 3 questões originais seguindo rigorosamente essas diretrizes, variando níveis cognitivos e gêneros textuais. Inclua respostas-modelo ou critérios de correção quando relevante. <i>(Adicione aqui qualquer especificação adicional, como foco em um eixo específico da BNCC ou restrição de formato.)</i>
Mistral	Para criar um prompt especializado que será usado para gerar questões educacionais de alta qualidade para alunos do 7º ano do Ensino Fundamental, com base na Base Nacional Comum Curricular (BNCC) de Língua Portuguesa e na Taxonomia de Bloom, podemos seguir as seguintes diretrizes: Diretrizes para o Prompt Contexto Significativo e Relevante: - As questões devem ser contextualizadas em situações do cotidiano dos alunos, como experiências pessoais, sociais, culturais e ambientais. - Utilizar temas transversais, como sustentabilidade, diversidade cultural, tecnologia, e cidadania. Enunciados Claros e Objetivos: - Os enunciados devem ser diretos e fáceis de entender, evitando ambiguidades. - Utilizar linguagem acessível, mas desafiadora, para estimular o pensamento crítico. Alinhamento com a BNCC: - As questões devem cobrir as competências e habilidades específicas da BNCC para o 7º ano, como leitura, escrita, oralidade, análise linguística e literatura. - Incluir habilidades de interpretação de textos, produção textual, gramática contextualizada, e análise de gêneros textuais. Taxonomia de Bloom: - As questões devem abranger diferentes níveis cognitivos da Taxonomia de Bloom, desde a lembrança e compreensão até a análise, avaliação e criação. - Exemplos de verbos para cada nível: - Lembrança: identificar, listar, memorizar. - Compreensão: explicar, resumir, interpretar. - Aplicação: aplicar, demonstrar, usar. - Análise: comparar, contrastar, diferenciar. - Avaliação: avaliar, justificar, criticar. - Criação: criar, projetar, propor. Formatos de Questões:

- As questões podem ser de múltipla escolha, dissertativas, verdadeiro ou falso, associação, preenchimento de lacunas, entre outros.
- Incluir questões que envolvam diferentes gêneros textuais, como contos, poemas, notícias, charges, e infográficos.

Exemplo de Prompt

"Crie questões educacionais de alta qualidade para alunos do 7º ano do Ensino Fundamental, alinhadas com a Base Nacional Comum Curricular (BNCC) de Língua Portuguesa e a Taxonomia de Bloom. As questões devem incluir contextos significativos e relevantes para estudantes desta faixa etária, com enunciados claros e objetivos. Utilize diferentes formatos de questões e aborde temas transversais, como sustentabilidade, diversidade cultural, tecnologia, e cidadania. As questões devem cobrir as competências e habilidades específicas da BNCC para o 7º ano, como leitura, escrita, oralidade, análise linguística e literatura. Inclua questões que abrangem diferentes níveis cognitivos da Taxonomia de Bloom, desde a lembrança e compreensão até a análise, avaliação e criação."

Exemplo de Questão

Contexto: Você está lendo um conto sobre um jovem que descobre um segredo sobre sua família. O conto aborda temas como identidade e pertencimento.

Enunciado: Após ler o conto, responda às seguintes questões:

1. **Lembrança:** Liste os personagens principais do conto e descreva brevemente cada um.
2. **Compreensão:** Explique o significado do segredo descoberto pelo protagonista e como isso afeta sua vida.
3. **Aplicação:** Imagine que você é o protagonista. Como você agiria ao descobrir o segredo? Justifique sua resposta.
4. **Análise:** Compare o relacionamento do protagonista com sua família antes e depois da descoberta do segredo.
5. **Avaliação:** Avalie a importância do tema "identidade" no conto. Como esse tema é desenvolvido ao longo da história?
6. **Criação:** Crie um final alternativo para o conto, mantendo os temas de identidade e pertencimento.

Este prompt e exemplo de questão podem ser adaptados e expandidos conforme necessário para atender às necessidades específicas dos alunos e do currículo.

Prompt Especializado para Geração de Questões - 7º Ano - Língua Portuguesa

Você é um(a) professor(a) experiente em Língua Portuguesa do Ensino Fundamental II, com conhecimento aprofundado da BNCC e da Taxonomia de Bloom. Sua tarefa é elaborar **questões educacionais de alta qualidade** para alunos do **7º ano do Ensino Fundamental**. Siga as diretrizes abaixo para gerar cada questão:

1. Objetivo da Questão

- A questão deve estar alinhada com **competências específicas da BNCC de Língua Portuguesa para o 7º ano**, por exemplo:
 - (EF69LP01) Analisar o efeito de sentido produzido pelo uso de diferentes tempos e modos verbais.
 - (EF69LP04) Analisar a função e o efeito de sentido de tempos verbais em textos narrativos.
 - (EF69LP10) Identificar o uso de recursos coesivos (conjunções, pronomes etc.) na construção de sentido do texto.
 - (EF69LP13) Avaliar a intenção e o efeito de sentido de recursos linguísticos em gêneros argumentativos.

2. Nível da Taxonomia de Bloom

Cada questão deve indicar e explorar um dos seguintes níveis:

- **Lembrar** (reconhecer, listar, definir)
- **Compreender** (interpretar, resumir, classificar)
- **Aplicar** (usar em situações novas)
- **Analisar** (comparar, categorizar, inferir)
- **Avaliar** (criticar, justificar, argumentar)
- **Criar** (elaborar, compor, reescrever)

3. Formato da Questão

- **Tipos permitidos:** múltipla escolha, verdadeiro/falso, resposta curta, associação, produção textual, reescrita, interpretação com imagem ou tirinha, etc.
- Incluir **um contexto significativo para adolescentes de 12 a 13 anos** (ex: redes sociais, meio ambiente, esportes, escola, amizades, tecnologia, cultura pop, etc.).
- O enunciado deve ser **claro, objetivo e acessível** à faixa etária.

4. Estrutura Esperada da Saída

markdown

Copiar código

Habilidade BNCC:

(EF69LPXX) — [Descrição da habilidade]

Nível da Taxonomia de Bloom:

[Ex: Analisar]

Tipo de Questão:

[Ex: Múltipla escolha]

Enunciado:

[Texto-base ou contexto, se necessário]

[Enunciado da questão propriamente dito]

Alternativas (se aplicável):

A)

B)

C)

D)

Gabarito:

[Letra correta ou resposta modelo]

Justificativa pedagógica:

[Explicação do que está sendo avaliado e por que a resposta correta é adequada]

Exemplo de Aplicação do Prompt

markdown

Copiar código

Habilidade BNCC:

(EF69LP04) — Analisar a função e o efeito de sentido de tempos verbais em textos narrativos.

Nível da Taxonomia de Bloom: Analisar

Tipo de Questão: Múltipla escolha

Enunciado: Leia o trecho a seguir retirado de uma história publicada em um blog:

"Quando Mariana entrou na sala, todos ficaram em silêncio. Ela não sabia o motivo, mas sentia que algo estava errado."

Com base nesse trecho, analise o uso dos tempos verbais.

O uso dos tempos verbais no trecho contribui para:

A) indicar uma ação futura com ênfase nas intenções da personagem.

B) descrever ações repetitivas que fazem parte da rotina dos personagens.

C) narrar fatos passados com foco nas emoções e reações dos personagens.

D) apresentar uma opinião pessoal da narradora sobre a personagem

	principal. ### Gabarito:C ### Justificativa pedagógica: A questão avalia a capacidade do aluno de reconhecer o efeito de sentido gerado pelo uso dos tempos verbais no passado, típico de textos narrativos. A alternativa correta mostra que o tempo verbal contribui para narrar acontecimentos e emoções passadas.
--	---

Tabela A.5 - Questões geradas com o Meta - Prompt.

LLMs	Questões criadas
Gemini	Com certeza! Como seu gerador de questões educacionais, vou criar três questões de Língua Portuguesa para o 7º ano do Ensino Fundamental, seguindo todas as suas diretrizes, focando em diferentes níveis da Taxonomia de Bloom e alinhadas à BNCC. Questão 1 • Nível da Taxonomia de Bloom: Compreender • Habilidade(s) da BNCC: EF69LP03 - Comparar textos e opiniões; EF67LP04 - Análise de gêneros. • Contexto/Texto Base: Imagem de um meme popular: Um "gif" de um gato olhando surpreso para a tela de um computador, com a legenda: "Eu depois de maratonar uma série inteira em um dia e perceber que a escola começa amanhã." • Enunciado da Questão: Observando o meme acima, qual é a principal emoção expressa pelo gato e qual situação do cotidiano de um estudante do 7º ano ele representa de forma bem-humorada? Explique com suas palavras. • Tipo de Questão: Dissertativa Curta • Gabarito/Sugestão de Resposta: A principal emoção expressa pelo gato é o espanto/surpresa, misturado talvez com um pouco de arrependimento ou cansaço. Ele representa a situação em que o estudante se dedicou intensamente a uma atividade de lazer (maratonar uma série) e, de repente, se dá conta de que o período de descanso acabou e as responsabilidades escolares estão prestes a começar novamente. O humor surge do contraste entre a diversão excessiva e a realidade iminente. Questão 2 • Nível da Taxonomia de Bloom: Aplicar

- **Habilidade(s) da BNCC:** EF07LP09 - Identificar classes de palavras; EF67LP07 - Produção de textos.
- **Contexto/Texto Base:**

Cenário: Maria está escrevendo um post para as redes sociais sobre a importância de combater o bullying na escola. Ela escreveu o seguinte rascunho:

"O **bullying** é um problema **sério** que afeta muitos adolescentes **diariamente**. Precisamos ser mais **gentis** uns com os outros e **sempre** defender quem precisa de ajuda. Juntos, podemos criar um ambiente escolar **melhor** para **todos**."

- **Enunciado da Questão:** No rascunho de Maria, identifique um **substantivo**, um **adjetivo** e um **advérbio**. Em seguida, reescreva a frase que contém o advérbio, substituindo-o por outro advérbio de mesma classificação, mas que mude o sentido da frase.
- **Tipo de Questão:** Dissertativa Curta
- **Gabarito/Sugestão de Resposta:**
 - **Substantivo:** bullying (ou problema, adolescentes, ajuda, ambiente, escola)
 - **Adjetivo:** sério (ou gentis, melhor, todos)
 - **Advérbio:** "diariamente" (advérbio de tempo) ou "sempre" (advérbio de tempo).
 - **Reescrita da frase com advérbio de tempo substituído:**
 - Original: "O bullying é um problema sério que afeta muitos adolescentes **diariamente**."
 - Substituição: "O bullying é um problema sério que afeta muitos adolescentes **ocasionalmente**." (o sentido da frase muda para indicar que o problema acontece de vez em quando, e não todos os dias).
 - Original: "Precisamos ser mais gentis uns com os outros e **sempre** defender quem precisa de ajuda."
 - Substituição: "Precisamos ser mais gentis uns com os outros e **raramente** defender quem precisa de ajuda." (o sentido da frase muda completamente, para o oposto da ideia original).

Questão 3

- **Nível da Taxonomia de Bloom:** Avaliar
- **Habilidade(s) da BNCC:** EF69LP08 - Opinião e argumentação; EF67LP06 - Análise crítica de mídias.
- **Contexto/Texto Base:**

Notícia (descrição): Seria apresentada uma notícia curta (ou trecho dela) de um site de notícias confiável, com o

	título: "Jovens usam redes sociais para promover campanhas de conscientização ambiental e social". O texto descreveria exemplos de como adolescentes estão utilizando plataformas como TikTok e Instagram para espalhar mensagens positivas sobre temas como consumo consciente, diversidade e combate à desinformação. • Enunciado da Questão: A notícia afirma que jovens estão usando as redes sociais para promover campanhas de conscientização ambiental e social. Você concorda que as redes sociais são ferramentas eficazes para que adolescentes se engajem em causas importantes? Justifique sua resposta apresentando, no mínimo, dois argumentos que apoiem ou refutem essa ideia, considerando os potenciais positivos e negativos dessas plataformas. • Tipo de Questão: Dissertativa Aberta • Gabarito/Sugestão de Resposta: O aluno pode concordar ou discordar, desde que apresente argumentos consistentes. ○ Argumentos a favor (exemplos): ▪ Amplo alcance e velocidade: As redes sociais permitem que as mensagens se espalhem rapidamente para um grande número de pessoas, alcançando jovens em diferentes lugares. ▪ Linguagem acessível e engajamento: O formato dos conteúdos (vídeos curtos, memes, posts com imagens) é atraente para o público jovem, facilitando o engajamento e a compreensão. ▪ Construção de comunidades: Os jovens podem se conectar com outras pessoas que compartilham dos mesmos interesses e causas, fortalecendo o movimento. ○ Argumentos contra (exemplos): ▪ Disseminação de desinformação: A facilidade de compartilhar conteúdo também abre espaço para informações falsas ou superficiais, que podem prejudicar a credibilidade das campanhas. ▪ "Engajamento de fachada": O simples "curtir" ou "compartilhar" pode não se traduzir em ações concretas no mundo real. ▪ Pressão e comparação social: O uso excessivo das redes sociais para esses fins pode gerar pressão para estar sempre engajado, além de comparações que podem ser prejudiciais à saúde mental.
DeepSeek	Aqui estão três questões originais alinhadas à BNCC (7º ano) e à Taxonomia de Bloom, com variedade de gêneros e níveis cognitivos:

Questão 1

Habilidade BNCC: EF69LP03 (Inferir o sentido de palavras/expressões em contextos específicos).

Nível Cognitivo: Compreender.

Gênero: Meme (texto multimodal).

Contexto: Meme com a imagem de um gato cansado e a frase: "Segunda-feira me pegou no 'tô sem condições'."

Pergunta:

A expressão "tô sem condições" no meme não indica falta de condições financeiras, mas um estado emocional. Que significado ela transmite no contexto? Justifique com base no suporte visual (imagem do gato).

Resposta-modelo:

A expressão transmite a ideia de cansaço ou desânimo, comum no início da semana. A imagem do gato exausto reforça esse sentido, criando humor ao associar o sentimento humano ao animal.

Critérios de correção:

- Identificar o sentido figurado (1 pt);
- Relacionar texto e imagem (1 pt).

Questão 2

Habilidade BNCC: EF07LP20 (Analisar a relação entre recursos verbais e não verbais em textos multissemióticos).

Nível Cognitivo: Analisar.

Gênero: Infográfico sobre reciclagem (com ícones e dados estatísticos).

Pergunta:

O infográfico mostra que 60% do lixo plástico não é reciclado. Como o uso da cor vermelha no gráfico e o ícone de um planeta "triste" contribuem para o efeito de sentido da mensagem?

Resposta-modelo:

A cor vermelha alerta para o perigo da poluição, enquanto o ícone do planeta triste humaniza o problema, apelando para a empatia do leitor e reforçando a urgência da reciclagem.

Critérios de correção:

- Explicar a função da cor (1 pt);
- Interpretar o ícone (1 pt).

Questão 3

Habilidade BNCC: EF69LP11 (Produzir narrativas ficcionais com desenvolvimento de enredo e personagens).

	Nível Cognitivo: Criar. Contexto: Tema "Um desafio inesperado na escola". Pergunta: Escreva um conto (15-20 linhas) sobre um(a) estudante que precisa resolver um problema na escola usando criatividade. Inclua: • Um conflito claro (ex.: um evento cancelado, um mal-entendido); • Um diálogo que revele a personalidade do personagem; • Um desfecho surpreendente. CrITÉRIOS de correção: • Presença dos elementos solicitados (2 pts); • Coesão e originalidade (1 pt); • Uso adequado da norma culta (1 pt). Diferenciais das questões: 1. Contextos atuais (memes, sustentabilidade) para engajar jovens. 2. Progressão cognitiva: da compreensão à criação. 3. Multimodalidade: integração de linguagens (visual, verbal). Adapte os temas conforme o interesse da turma!
Mistral	Vamos criar um exemplo de questão educacional seguindo as diretrizes fornecidas. Este exemplo será alinhado com a Base Nacional Comum Curricular (BNCC) para o 7º ano do Ensino Fundamental e utilizará a Taxonomia de Bloom para abranger diferentes níveis cognitivos. Exemplo de Questão Contexto: Você está lendo uma notícia sobre um projeto de sustentabilidade implementado em uma escola da sua cidade. O projeto envolve a criação de uma horta comunitária e a reciclagem de materiais. A notícia aborda temas como sustentabilidade, trabalho em equipe e impacto ambiental. Enunciado: Após ler a notícia, responda às seguintes questões: 1. Lembrança: Identifique os principais objetivos do projeto de sustentabilidade mencionado na notícia. 2. Compreensão: Explique como a horta comunitária contribui para a sustentabilidade e para a comunidade escolar. 3. Aplicação: Se você pudesse participar desse projeto, que ações específicas você proporia para melhorar ainda mais a sustentabilidade na sua escola? Justifique suas sugestões. 4. Análise: Compare os benefícios da reciclagem de materiais com os da criação de uma horta comunitária. Quais são as semelhanças e diferenças entre essas duas iniciativas?

	5. Avaliação: Avalie a importância do trabalho em equipe para o sucesso do projeto de sustentabilidade. Como o trabalho em equipe é destacado na notícia? 6. Criação: Crie um plano de ação para implementar um novo projeto de sustentabilidade na sua escola, diferente do mencionado na notícia. Descreva as etapas do projeto e como ele poderia ser executado. Adaptação e Expansão Este exemplo de questão pode ser adaptado para incluir outros temas transversais e diferentes gêneros textuais, conforme necessário. Além disso, as questões podem ser modificadas para atender a diferentes habilidades e competências específicas da BNCC, garantindo que os alunos sejam desafiados de maneira adequada e estimulante.
ChatGPT	Claro! Abaixo seguem 3 questões elaboradas com base nas diretrizes fornecidas, cada uma focando em uma habilidade da BNCC e um nível da Taxonomia de Bloom, com temas relevantes para adolescentes do 7º ano: Questão 1 ### Habilidade BNCC: (EF69LP01) — Analisar o efeito de sentido produzido pelo uso de diferentes tempos e modos verbais. ### Nível da Taxonomia de Bloom: Analisar ### Tipo de Questão: Múltipla escolha ### Enunciado: Leia o trecho da crônica a seguir: > _"Se eu estudasse mais um pouco, talvez fosse melhor na prova de matemática. Mas agora já passou, e eu só posso esperar pelo próximo teste."_ A análise dos tempos verbais utilizados no trecho permite concluir que o narrador: ### Alternativas: A) Está descrevendo fatos do presente com foco em ações rotineiras. B) Expressa um desejo futuro com certeza de que vai realizá-lo. C) Relata uma hipótese do passado e uma constatação presente. D) Narra uma história totalmente ficcional com verbos no pretérito. ### Gabarito: C

Justificativa pedagógica:

A questão explora o uso de tempos e modos verbais para expressar hipóteses (condicional com “se eu estudasse”) e conclusões no presente. O aluno precisa ****analisar**** como os tempos verbais são empregados para construir o sentido da narrativa pessoal e reflexiva.

Questão2

Habilidade BNCC:

(EF69LP10) — Identificar o uso de recursos coesivos (conjunções, pronomes etc.) na construção de sentido do texto.

Nível da Taxonomia de Bloom:

Compreender

Tipo de Questão: Verdadeiro ou falso

Enunciado:

Leia o trecho abaixo de uma postagem em um blog estudantil:

> _"Estávamos prontos para começar a apresentação. No entanto, o projetor não funcionava, e tivemos que improvisar com cartazes."_

Sobre o uso do conectivo “no entanto”, assinale (V) para verdadeiro ou (F) para falso:

- () O conectivo indica oposição entre as ideias.
- () “No entanto” pode ser substituído por “então” sem alterar o sentido.
- () A frase mostra que algo inesperado aconteceu.
- () O uso de “no entanto” reforça uma explicação.

Gabarito: V – F – V – F

Justificativa pedagógica:

A questão avalia a ****compreensão**** do aluno sobre o papel de conectivos no texto, especialmente aqueles que expressam ****oposição ou contraste****. O uso de “no entanto” mostra que houve um imprevisto, contrariando a expectativa criada na primeira oração.

Questão 3

Habilidade BNCC:

(EF69LP13) — Avaliar a intenção e o efeito de sentido de recursos linguísticos em gêneros argumentativos.

Nível da Taxonomia de Bloom: Avaliar

Tipo de Questão: Resposta curta

Enunciado: Leia o trecho de uma carta de leitor publicada em uma revista jovem:

	> _"Não faz sentido dizer que videogames tornam os jovens violentos. Muitos estudos já mostraram que o comportamento agressivo está relacionado a outros fatores, como o ambiente familiar e a saúde mental."_ Com base no trecho acima, explique com suas palavras o efeito de sentido provocado pelo uso da expressão "não faz sentido dizer que...". ### Gabarito: O autor busca **desqualificar um argumento comum** com base em uma opinião forte, para **introduzir e justificar seu ponto de vista** com embasamento. A expressão "não faz sentido" é usada para **refutar uma ideia** e preparar o leitor para aceitar a tese defendida. ### Justificativa pedagógica: A atividade propõe que o aluno **avalie** o uso de uma expressão argumentativa em um gênero de opinião. O objetivo é perceber como o autor organiza a argumentação e convence o leitor, exercitando a leitura crítica.
--	--

Apêndice B – Questões criadas e Prompts usados para o experimento

Tabela B.1 - Questões geradas por um professor.

Professor	Questões criadas
	As dores do mundo Elias José “Sinto bem fundo todas as dores do mundo. Só que meu poema não conseguiu tocar em feridas maiores, Abro os jornais e leio e choro e me arrepio com a fome com a guerra. com a aids, com a violência, com a desnutrição do verde e da vida. Tento escrever, mas sai um poema impotente. Fico pensando: as dores do mundo pedem canções ou exigem ação?” JOSÉ, E. Cantigas de adolescer. São Paulo: Atual, 1992. Habilidade EF69LP48: Interpretar efeitos produzidos pelo uso de recursos expressivos sonoros (rimas, aliterações, etc.), semânticos (figuras de linguagem), gráfico-espaciais (distribuição da mancha gráfica) e imagens em poemas, e sua relação com o texto verbal. <u>Questão nº01</u> Chamamos de eu poético a voz que se manifesta e expressa os sentimentos do narrador no poema. Dessa forma, após a leitura do poema de Elias José, transcreva algumas dores do mundo citadas pelo eu poético. <u>Questão nº02</u> Já aprendemos que PERSONIFICAÇÃO é recurso da linguagem figurada que consiste em atribuir qualidades humanas a seres não humanas.

Identifique e copie um verso em que temos personificação.

Habilidade EF07LP11:

Identificar, em textos lidos ou de produção própria, períodos compostos nos quais duas orações são conectadas por vírgula, ou por conjunções que expressem soma de sentido (conjunção "e") ou oposição de sentidos (conjunções "mas", "porém").

Questão nº03

Leia os seguintes versos:

"Tento escrever,
mas sai um poema impotente."

No poema, aparece a palavra MAS, uma conjunção que indica a ideia de oposição, adversidade. Que outra conjunção poderia substituí-la sem prejuízo de sentido?

Tabela B.2 - Questões geradas com o Prompt.

LLMs	Questões criadas
DeepSeek	Aqui estão três questões alinhadas à BNCC e à Taxonomia de Bloom, com contextos relevantes para alunos do 7º ano: 1. Nível Lembrar (Taxonomia de Bloom) – BNCC: EF67LP03 (Morfologia) Contexto: Você está ajudando um colega a estudar para uma prova e ele pergunta: <i>"Quais são as classes gramaticais das palavras destacadas na frase abaixo?"</i> Frase: "Ontem, comprei dois livros incríveis na feira e li um rapidamente." a) dois – numeral / incríveis – adjetivo / feira – substantivo / li – verbo / rapidamente – advérbio b) dois – adjetivo / incríveis – substantivo / feira – verbo / li – advérbio / rapidamente – numeral c) dois – advérbio / incríveis – verbo / feira – substantivo / li – adjetivo / rapidamente – numeral Resposta esperada: Letra a. 2. Nível Aplicar (Taxonomia de Bloom) – BNCC: EF69LP08 (Produção textual) Contexto: Você está escrevendo um post para o blog da escola sobre <i>"Como economizar água no dia a dia"</i>. Reescreva a frase abaixo, substituindo a linguagem informal por uma variante culta formal:

	Frase informal: "Você pode fechar a torneira quando estiver escovando os dentes pra evitar desperdício." Resposta esperada: "É recomendável fechar a torneira enquanto escova os dentes para evitar desperdício." 3. Nível Analisar (Taxonomia de Bloom) – BNCC: EF67LP24 (Interpretação textual) Contexto: Leia o trecho do conto "O Espelho", de Machado de Assis, e responda: <i>"O meu primeiro retrato foi pintado aos dezessete anos. Era um belo retrato, parecia-me até mais bonito do que eu..."</i> Pergunta: Qual é o tom predominante no trecho e como ele influencia a compreensão do narrador? a) Tom humorístico – mostra que o narrador não leva a si mesmo a sério. b) Tom nostálgico – revela uma reflexão sobre o passado e a autoimagem. c) Tom irônico – sugere uma crítica à vaidade humana. Resposta esperada: Letra b (nostalgia e reflexão sobre a autoimagem). Justificativa: • A primeira questão avalia conhecimento básico (Lembrar), essencial para análise morfológica. • A segunda exige aplicação prática de normas cultas (Aplicar), útil em produções textuais. • A terceira trabalha análise crítica (Analisar), incentivando a interpretação de nuances textuais. Todas seguem a BNCC e envolvem contextos do cotidiano ou cultura literária, adequados ao 7º ano.
Mistral	Aqui estão três questões sobre Língua Portuguesa para alunos do 7º ano do Ensino Fundamental, seguindo a BNCC e os diferentes níveis da Taxonomia de Bloom: Questão 1: Compreensão (Nível de Compreensão da Taxonomia de Bloom) Contexto: Você está lendo um trecho de um livro de aventuras. O personagem principal encontra um mapa do tesouro com as seguintes instruções:

	<i>"Caminhe vinte passos ao norte da grande árvore, depois quinze passos a leste. Você encontrará uma pedra redonda. Sob ela está o tesouro."</i> Enunciado: Explique, com suas próprias palavras, as instruções do mapa do tesouro. Questão 2: Aplicação (Nível de Aplicação da Taxonomia de Bloom) Contexto: Você está escrevendo um e-mail para um amigo que se mudou para outra cidade. Você quer contar sobre um evento especial que aconteceu na escola. Enunciado: Escreva um e-mail para o seu amigo descrevendo o evento especial que aconteceu na escola. Inclua detalhes sobre o que aconteceu, quem participou e como você se sentiu. Questão 3: Análise (Nível de Análise da Taxonomia de Bloom) Contexto: Você leu um conto sobre um menino que superou seus medos. O conto termina com a seguinte frase: <i>"Ele percebeu que o verdadeiro tesouro não era o ouro, mas a coragem que encontrou dentro de si."</i> Enunciado: Analise o significado da frase final do conto. O que o autor quis dizer com "o verdadeiro tesouro não era o ouro, mas a coragem que encontrou dentro de si"? Dê exemplos do conto que apoiem sua análise. Essas questões abordam diferentes habilidades cognitivas e estão alinhadas com os objetivos de aprendizagem da BNCC para o 7º ano do Ensino Fundamental.
ChatGPT	Claro! Abaixo estão três questões de Língua Portuguesa para o 7º ano do Ensino Fundamental, alinhadas à BNCC e aos diferentes níveis da Taxonomia de Bloom, com contextos significativos para a faixa etária. Cada questão apresenta a habilidade da BNCC correspondente, o nível de Bloom e uma breve justificativa pedagógica.  Questão 1 – Compreensão (Nível: Compreender) BNCC: EF67LP28 – Ler, de forma autônoma, e compreender textos de gêneros variados. Contexto: Leia o texto abaixo, retirado de uma reportagem sobre redes sociais e adolescentes: "Geração Conectada: O Impacto das Redes Sociais na Vida dos Jovens"

Aos 13 anos, Marina passa cerca de 4 horas por dia navegando pelas redes sociais. Entre stories no Instagram, vídeos no TikTok e conversas no WhatsApp, ela divide seu tempo livre entre o mundo virtual e os estudos. "É difícil desconectar", admite a estudante. "Parece que sempre tem algo novo acontecendo."

Segundo especialistas, o uso excessivo das redes sociais pode afetar o sono, o rendimento escolar e até mesmo a autoestima dos jovens. Dr. Carlos Mendes, psicólogo especialista em comportamento adolescente, explica: "É importante estabelecer limites saudáveis para o uso da tecnologia, especialmente nessa fase de desenvolvimento."

Marina reconhece que precisa encontrar um equilíbrio: "Estou tentando deixar o celular longe na hora de estudar e dormir pelo menos uma hora mais cedo."

Com base no texto, responda:

- a) Qual é o problema central apresentado na reportagem?
- b) Identifique duas consequências do uso excessivo das redes sociais mencionadas no texto.
- c) Que solução Marina encontrou para lidar com a situação?

Questão 2 – Aplicação (Nível: Aplicar)

BNCC: EF67LP10 – Identificar e analisar o uso de elementos coesivos (pronomes, advérbios, conjunções, etc.) em textos diversos.

Enunciado: Observe este trecho de um blog escrito por uma estudante:

"Eu fui ao parque com minha amiga Ana. Nós brincamos bastante no escorregador. Depois, **ela** quis tomar um sorvete, então fomos até a sorveteria."

Pergunta: O pronome destacado ("ela") refere-se a quem?

Alternativas:

- A) À narradora do texto.
- B) Ao sorveteiro.
- C) À amiga Ana.
- D) À mãe da narradora.

Resposta correta: C

Justificativa: A questão avalia a capacidade de o aluno aplicar seus conhecimentos sobre coesão referencial, habilidade essencial para a produção textual e compreensão.

	✳ Questão 3 – Criação (Nível: Criar) BNCC: EF67LP28 – Produzir, revisar e editar textos de diferentes gêneros com coerência, coesão e adequação ao contexto. Enunciado: Imagine que sua escola vai criar um jornal digital, e você foi convidado(a) a escrever uma pequena reportagem sobre um evento escolar recente. Tarefa: Escreva uma reportagem de até 10 linhas, contando o que aconteceu no evento, quem participou e por que ele foi importante para a escola. Critérios de avaliação: • Clareza na apresentação das informações • Uso adequado de linguagem jornalística (fatos, datas, participantes) • Coesão e coerência textual Justificativa: Esta proposta estimula a criatividade e a produção de texto em um gênero próximo do cotidiano dos estudantes, promovendo competências de autoria e domínio da linguagem formal.
--	--

Tabela B.3 - Meta – Prompt para a criação das questões.

LLMs	Questões criadas
DeepSeek	Prompt para Gerar Questões Alinhadas à BNCC e à Taxonomia de Bloom – 7º Ano (Língua Portuguesa) Instruções: Crie 3 questões para alunos do 7º ano do Ensino Fundamental, alinhadas às habilidades da BNCC de Língua Portuguesa (ex: EF67LPxx, EF69LPxx) e articuladas com os níveis cognitivos da Taxonomia de Bloom (lembrar, compreender, aplicar, analisar, avaliar, criar). As questões devem: Contextualização significativa: Usar temas relevantes para adolescentes (redes sociais, sustentabilidade, cultura jovem, games, etc.). Formato diversificado: Podem ser abertas, fechadas (múltipla escolha com 4 opções), ou associativas.

Objetividade: Enunciados claros, com comando direto (ex: "Analise", "Justifique", "Reescreva").

Formatação pronta para prova: Fonte Arial 11, espaço 1,5, com indicativo de valor (ex: "Valor: 2,0 pontos").

Exemplo de Saída Esperada:

QUESTÃO 1 (Valor: 2,0 pontos)

Habilidade BNCC: EF67LP16 – Inferir o efeito de sentido produzido pelo uso de recursos expressivos em textos.

Taxonomia de Bloom: Analisar

Leia o trecho da música "Exagerado" (Cazuza):

"Eu sei que eu sempre fui exagerado / Eu gosto de amar além da conta"

a) Que efeito de sentido o uso da palavra "além da conta" cria no texto? Justifique sua resposta com elementos do trecho.

QUESTÃO 2 (Valor: 1,5 pontos)

Habilidade BNCC: EF69LP04 – Produzir resumos com síntese das ideias principais de um texto.

Taxonomia de Bloom: Aplicar

Leia o post abaixo de um perfil sobre sustentabilidade:

"Você sabia que 1 canudo plástico leva 400 anos para se decompor? Troque por opções reutilizáveis!"

a) Reescreva o post em uma única frase, mantendo a ideia central e o caráter persuasivo.

QUESTÃO 3 (Valor: 2,5 pontos)

Habilidade BNCC: EF67LP24 – Identificar marcas de variação linguística em diferentes situações de comunicação.

Taxonomia de Bloom: Avaliar

Compare as duas frases:

"Mano, a gente vai chegar lá rapidão!"

"Senhor, chegaremos ao local em breve."

a) Em qual situação cada frase seria mais adequada? Explique como a variação linguística impacta a comunicação.

	Observações: Inclua gabarito comentado se as questões forem objetivas. Priorize verbos operacionais (identifique, discuta, proponha) para direcionar a ação do aluno. Use textos reais (trechos de livros, posts, letras de música) para ancorar as questões. Adapte conforme a unidade temática trabalhada (leitura, escrita, oralidade ou análise linguística).
Mistral	Crie três questões educacionais de alta qualidade para alunos do 7º ano do Ensino Fundamental, seguindo as diretrizes da Base Nacional Comum Curricular (BNCC) de Língua Portuguesa e aplicando os conceitos da Taxonomia de Bloom. As questões devem incluir contextos significativos e relevantes para estudantes desta faixa etária, com enunciados claros e objetivos. As questões podem ser de múltipla escolha, dissertativas ou de verdadeiro ou falso, e devem estar formatadas para serem aplicadas diretamente em um documento de prova. Questão 1: Compreensão de Texto (Nível de Compreensão da Taxonomia de Bloom) Texto: "Na escola, João sempre foi um aluno curioso. Ele adorava ler livros de aventura e ciência. Um dia, sua professora de português pediu para que a turma escrevesse uma redação sobre um herói da vida real. João decidiu escrever sobre Santos Dumont, o pai da aviação. Ele passou horas pesquisando e escrevendo, e no final, sua redação foi a mais elogiada da turma." Enunciado: Com base no texto acima, responda: Qual foi o tema da redação que João escreveu e por que ele escolheu esse tema? Formato da Questão: Dissertativa Questão 2: Análise Linguística (Nível de Análise da Taxonomia de Bloom) Texto: "O meio ambiente é um tema muito importante nos dias de hoje. A preservação das florestas, a economia de água e a reciclagem de lixo são ações que podem fazer a diferença. Cada um de nós pode contribuir para um planeta mais saudável."

	Enunciado: No texto acima, identifique e sublinhe os substantivos relacionados ao meio ambiente. Em seguida, classifique-os em comuns ou próprios. Formato da Questão: Dissertativa Questão 3: Aplicação de Conhecimento (Nível de Aplicação da Taxonomia de Bloom) Enunciado: Imagine que você é um repórter e precisa escrever uma notícia sobre um evento importante que aconteceu em sua escola. Utilize os elementos básicos de uma notícia (título, lead e corpo do texto) para criar um texto informativo e atraente. Formato da Questão: Dissertativa Espero que este prompt atenda às suas necessidades! Se precisar de mais alguma coisa, é só avisar.
ChatGPT	Claro! Abaixo está um prompt cuidadosamente elaborado para gerar 3 questões educacionais de alta qualidade para alunos do 7º ano do Ensino Fundamental, alinhado com a BNCC de Língua Portuguesa e aplicando os níveis da Taxonomia de Bloom. O prompt pode ser usado com modelos de linguagem como o ChatGPT, Gemini, Claude, entre outros.  Prompt para geração de questões educacionais Crie 3 questões de Língua Portuguesa destinadas a alunos do 7º ano do Ensino Fundamental, conforme os seguintes critérios:  Taxonomia de Bloom As questões devem contemplar diferentes níveis da Taxonomia de Bloom: Lembrar/Compreender – uma questão que avalie a compreensão de conteúdos básicos, como identificar elementos de um texto. Aplicar/Analisar – uma questão que exija interpretação ou análise de um texto, imagem ou situação comunicativa. Avaliar/Criar – uma questão que leve o aluno a formular uma opinião crítica, criar uma continuação de texto, reescrever com outro ponto de vista, etc.  Base Nacional Comum Curricular (BNCC) As questões devem estar alinhadas aos seguintes descritores da BNCC para o 7º ano:

EF67LP01: Localizar informações explícitas em textos.

EF67LP05: Analisar efeitos de sentido decorrentes do uso de pontuação, tempos verbais, modos de organização do texto.

EF67LP12: Produzir textos coerentes e coesos, adequados à situação comunicativa.

 Contexto Significativo

Use temas e situações relevantes para adolescentes de 12 a 13 anos, como:

Vida escolar

Relações familiares ou de amizade

Tecnologia e redes sociais

Curiosidades científicas

Problemas do cotidiano (meio ambiente, bullying, etc.)

 Formato de Apresentação das Questões

Cada questão deve conter:

Um enunciado claro e objetivo

Um texto de apoio, se necessário (curto e acessível)

Instruções específicas (caso seja dissertativa, múltipla escolha, verdadeiro ou falso, etc.)

Gabarito ao final do material

 Exemplo de saída esperada (modelo de formato final)

Questão 1 — Nível: Compreensão (BNCC: EF67LP01)

Leia o trecho abaixo:

"Na volta às aulas, Júlia percebeu que sua melhor amiga havia mudado de escola. Mesmo triste, ela decidiu escrever uma carta para contar as novidades."

Pergunta:

Qual foi a primeira reação de Júlia ao perceber a ausência de sua amiga?

	() Ficou aliviada () Ficou triste () Ficou com raiva () Ficou animada Gabarito: (X) Ficou triste Questão 2 — Nível: Análise (BNCC: EF67LP05) Leia o trecho abaixo: "Na sala, reinava um silêncio tão intenso que até o som do relógio parecia ensurdecedor." Pergunta: Explique o efeito de sentido provocado pela expressão "o som do relógio parecia ensurdecedor". Resposta esperada: A expressão cria um efeito de exagero (hipérbole) para mostrar que o silêncio era tão grande que qualquer ruído parecia muito alto. Questão 3 — Nível: Criação (BNCC: EF67LP12) Imagine que você é Júlia, a personagem da primeira questão. Escreva uma resposta à carta da amiga que se mudou. Fale sobre seus sentimentos e conte uma novidade da escola. Use linguagem adequada e organize suas ideias de forma clara e coerente. Instrução: Sua resposta deve ter entre 8 e 12 linhas. Gabarito: Resposta dissertativa. Avaliar uso adequado da norma padrão, coesão e coerência, organização de ideias, e pertinência ao tema.
--	---

Tabela B.4 - Questões geradas com o Meta - Prompt.

LLMs	Questões criadas
DeepSeek	Aqui estão três questões alinhadas à BNCC e à Taxonomia de Bloom para o 7º ano de Língua Portuguesa, com temas relevantes para adolescentes e formatos diversificados: QUESTÃO 1 (Valor: 2,0 pontos) Habilidade BNCC: EF67LP16 – Inferir o efeito de sentido produzido pelo uso de recursos expressivos em textos. Taxonomia de Bloom: Analisar Tema: Redes sociais Leia o tweet de um influenciador digital:

"O algoritmo das redes sociais é um lobo em pele de cordeiro: parece inofensivo, mas devora seu tempo sem você perceber."

a) Que efeito de sentido o autor cria ao comparar o algoritmo a um "lobo em pele de cordeiro"? Justifique com elementos do texto.

Gabarito comentado (se necessário):

A metáfora sugere que o algoritmo, aparentemente útil (como um cordeiro), esconde um efeito nocivo (como um lobo), criticando seu impacto disfarçado na produtividade. O aluno deve identificar a dualidade "inofensivo/devorar" para justificar.

QUESTÃO 2 (Valor: 1,5 pontos)

Habilidade BNCC: EF69LP04 – Produzir resumos com síntese das ideias principais de um texto.

Taxonomia de Bloom: Aplicar

Tema: Sustentabilidade

Leia o texto de uma campanha escolar:

"Descarte irregular de lixo eletrônico contamina o solo e a água. Traga seus celulares e tablets velhos para a reciclagem no pátio da escola!"

a) Reescreva a mensagem em uma única frase, mantendo a ideia central e o apelo à ação.

Exemplo de resposta: "Recicle seus eletrônicos usados na escola e evite a contaminação do meio ambiente!"

QUESTÃO 3 (Valor: 2,5 pontos)

Habilidade BNCC: EF67LP24 – Identificar marcas de variação linguística em diferentes situações de comunicação.

Taxonomia de Bloom: Avaliar

Tema: Cultura jovem

Compare as falas: "Família, o rolê foi épico hoje!"

"Senhora, o passeio no shopping foi muito agradável."

a) Em qual contexto cada variante é adequada? Explique como a escolha linguística reflete a relação entre os interlocutores.

Gabarito comentado: A primeira é informal, típica de conversas entre amigos (marcas: "família", "rolê", "épico").

A segunda é formal, usada com adultos ou em situações que exigem polidez (marcas: "senhora", "passeio", "agradável").

	O aluno deve relacionar as escolhas lexicais ao nível de formalidade. Observações finais: Verbos operacionais: "Justifique", "Reescreva", "Explique" direcionam a ação. Textos reais: Tweet, campanha escolar e diálogos cotidianos. Adaptação: Pode ser ajustada para unidades temáticas específicas (ex.: substituir o tweet por um meme para análise multimodal). Se precisar de variações (como questões fechadas), posso adaptar!
Mistral	Aqui estão três questões educacionais de alta qualidade para alunos do 7º ano do Ensino Fundamental, seguindo as diretrizes da BNCC de Língua Portuguesa e aplicando os conceitos da Taxonomia de Bloom: Questão 1: Compreensão de Texto (Nível de Compreensão da Taxonomia de Bloom) Texto: "Na escola, João sempre foi um aluno curioso. Ele adorava ler livros de aventura e ciência. Um dia, sua professora de português pediu para que a turma escrevesse uma redação sobre um herói da vida real. João decidiu escrever sobre Santos Dumont, o pai da aviação. Ele passou horas pesquisando e escrevendo, e no final, sua redação foi a mais elogiada da turma." Enunciado: Com base no texto acima, responda: Qual foi o tema da redação que João escreveu e por que ele escolheu esse tema? Formato da Questão: Dissertativa Questão 2: Análise Linguística (Nível de Análise da Taxonomia de Bloom) Texto: "O meio ambiente é um tema muito importante nos dias de hoje. A preservação das florestas, a economia de água e a reciclagem de lixo são ações que podem fazer a diferença. Cada um de nós pode contribuir para um planeta mais saudável." Enunciado: No texto acima, identifique e sublinhe os substantivos relacionados ao meio ambiente. Em seguida, classifique-os em comuns ou próprios. Formato da Questão: Dissertativa

	Questão 3: Aplicação de Conhecimento (Nível de Aplicação da Taxonomia de Bloom) Enunciado: Imagine que você é um repórter e precisa escrever uma notícia sobre um evento importante que aconteceu em sua escola. Utilize os elementos básicos de uma notícia (título, lead e corpo do texto) para criar um texto informativo e atraente. Formato da Questão: Dissertativa Essas questões foram elaboradas para serem aplicadas diretamente em um documento de prova e visam desenvolver diferentes habilidades cognitivas dos alunos, conforme a Taxonomia de Bloom.
ChatGPT	Claro! Abaixo estão 3 questões de Língua Portuguesa para o 7º ano do Ensino Fundamental, elaboradas conforme os critérios indicados: Taxonomia de Bloom, BNCC, contexto significativo e formato esperado.  Questão 1 — Nível: Compreensão (BNCC: EF67LP01) Leia o trecho a seguir: “Durante o intervalo, Luana e seus colegas decidiram organizar uma campanha de arrecadação de alimentos para ajudar famílias carentes da comunidade. A ideia foi bem recebida pela direção da escola.” Pergunta: Qual foi a iniciativa tomada por Luana e seus colegas? () Fizeram uma festa na escola () Organizaram uma campanha de doação () Realizaram uma excursão cultural () Criaram um grupo de estudos Gabarito: (X) Organizaram uma campanha de doação  Questão 2 — Nível: Análise (BNCC: EF67LP05) Leia a frase abaixo, dita por um dos personagens durante uma conversa por mensagem: “Se continuar assim, esse grupo vai virar uma bomba-relógio prestes a explodir!” Pergunta: O que o personagem quis expressar com a metáfora “bomba-relógio prestes a explodir”? Instrução: Explique o sentido figurado da expressão e o que ela revela sobre a situação do grupo.

Resposta esperada:

A expressão indica que há tensão ou conflitos crescentes no grupo, e que uma discussão ou briga pode acontecer a qualquer momento. É um uso de linguagem figurada para mostrar perigo ou instabilidade.

** Questão 3 — Nível: Criação (BNCC: EF67LP12)**

Imagine a seguinte situação: Você viu uma postagem nas redes sociais sobre o aumento do bullying nas escolas e ficou sensibilizado(a). Decidiu escrever um pequeno texto de opinião para o mural da escola.

Instrução: Escreva um parágrafo opinativo com entre 8 e 10 linhas, em que você:

Expresse sua opinião sobre o bullying;

Apresente pelo menos uma sugestão para combater esse problema;

Use argumentos coerentes e linguagem adequada.

Gabarito:

Resposta dissertativa. Avaliação baseada em:

Clareza e pertinência da opinião;

Argumentação consistente;

Coesão e coerência textual;

Adequação à norma padrão da língua escrita.

Apêndice C – Termo de Consentimento Livre e Esclarecido (TCLE) e Questionário

Termo de Consentimento Livre e Esclarecido (TCLE)

PÚBLICO DA PESQUISA: Professores de Língua Portuguesa.

Leia o termo abaixo e, caso concorde com ele, prossiga com suas respostas à pesquisa.

1 Título da Pesquisa:

Estudo sobre o uso de *Meta-Prompting* em LLMs para a Elaboração de Exercícios: Na Perspectiva de Docentes

2 Pesquisador Principal:

Wendy Silva Cataldo

3 Instituição:

Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

4 Introdução:

Você está sendo convidado(a) a participar de uma pesquisa que visa investigar o impacto das técnicas de *Prompting* e *Meta-Prompting* aplicadas a modelos de linguagem de grande porte (LLMs) no processo de elaboração de exercícios para a disciplina de Língua Portuguesa. Em um cenário de crescente utilização da Inteligência Artificial na educação, este estudo propõe uma análise crítica das potencialidades desses modelos combinados a estratégia de engenharia de *prompts*, com o intuito de promover a geração de questões coerentes e alinhadas aos objetivos pedagógicos.

5 Objetivo da Pesquisa:

O objetivo desta pesquisa é analisar a qualidade das questões de Língua Portuguesa geradas por LLMs sob a perspectiva de docentes experientes. Avaliaremos a clareza, a relevância, a cobertura curricular, a gramaticalidade, a utilidade pedagógica e a capacidade de resposta das questões geradas.

6 Procedimentos da Pesquisa:

O objetivo desta pesquisa é analisar a qualidade das questões de Língua Portuguesa geradas por LLMs e por um professor especialista, sob a perspectiva de docentes experientes. Avaliaremos a clareza, a relevância, a cobertura curricular, a gramaticalidade, a utilidade pedagógica e a capacidade de resposta das questões geradas.

7 Benefícios:

Participar desta pesquisa permitirá uma melhor compreensão da qualidade das questões geradas por LLMs e como elas se comparam às elaboradas por professores. Espera-se que os resultados possam contribuir para a formulação de melhores práticas no uso dessas ferramentas na educação.

8 Confidencialidade:

Todas as informações coletadas serão tratadas com estrita confidencialidade e serão utilizadas exclusivamente para fins acadêmicos. Os dados serão armazenados de forma segura e apenas a equipe de pesquisa terá acesso a eles. Nenhuma informação pessoal identificável será divulgada.

9 Autonomia e Direitos do Participante:

Sua participação é completamente voluntária. Você tem o direito de recusar ou interromper sua participação a qualquer momento, sem qualquer prejuízo. Em caso de retirada, seus dados serão excluídos. Você também poderá esclarecer dúvidas a qualquer momento.

10 Contato dos Pesquisadores:

Em caso de dúvidas ou necessidade de informações adicionais, entre em contato com:

- Carlos Eduardo De Mello (pesquisador principal, orientador da pesquisa)
E-mail: mello@uniriotec.br
- Wendy Silva Cataldo (pesquisadora, mestranda)
E-mail: wendy.cataldo@edu.unirio.br
- Laura de Oliveira Fernandes Moraes (pesquisadora, coorientadora da pesquisa)
E-mail: laura@uniriotec.br

11 Consentimento:

Após ser informado(a) detalhadamente sobre os objetivos, procedimentos, benefícios e seus direitos, e tendo a oportunidade de fazer todas as perguntas que desejar, você pode concordar em participar desta pesquisa avançando neste formulário online. Ao avançar para responder às questões, você declara que foi devidamente informado(a) sobre a pesquisa "Estudo sobre o uso de Meta-Prompting em LLMs para a Elaboração de Exercícios: Na Perspectiva de Docentes", que teve a oportunidade de esclarecer suas dúvidas e que está ciente de que sua participação é voluntária e pode ser retirada a qualquer momento.

Obrigado por participar desta pesquisa!

- ☐ Li e concordo em participar da pesquisa.
- ☐ Li, mas não concordo em participar da pesquisa.

Questionário

Obrigado por participar desta pesquisa!

Seu feedback é fundamental para avaliar o impacto da ferramenta na criação das questões.

Esta pesquisa visa analisar a percepção dos professores sobre o uso de ferramentas de LLMs na criação de exercícios. Todas as suas respostas serão confidenciais e utilizadas apenas para fins de pesquisa.

1. Formação e Experiência Profissional

1 Qual sua idade?

- ☐ 20-29 anos
- ☐ 30-39 anos
- ☐ 40-49 anos

☐ 50-59 anos

☐ 60 anos ou mais

2 Você possui pós-graduação?

☐ Não

☐ Sim, Especialização.

☐ Sim, Mestrado.

☐ Sim, Doutorado.

3 Instituição de ensino que leciona atualmente (ou a principal, caso leccione em mais de uma)?

☐ Escola pública municipal

☐ Escola pública estadual

☐ Escola particular

☐ Instituto Federal de Educação, Ciência e Tecnologia

☐ Colégio Militar

☐ Outra: _____

☐ Não estou lecionando no momento

4 Há quanto tempo você atua como professor(a) (considerando todas as disciplinas e níveis)?

☐ Menos de 1 ano

☐ 1 a 5 anos

☐ 6 a 10 anos

☐ 11 a 15 anos

☐ 16 a 20 anos

☐ Mais de 20 anos

5 Com que frequência você elabora suas próprias questões para as avaliações que aplica?

☐ Sempre (100% das questões)

☐ Frequentemente (Mais da metade das questões)

☐ Ocasionalmente (Cerca de metade das questões)

☐ Raramente (Menos da metade das questões)

☐ Nunca (Utilizo apenas questões de bancos, livros, etc.)

6 Em média, quanto tempo você dedica à criação de exercícios e avaliações para uma turma por semana?

☐ Menos de 1 hora

☐ 1 a 2 horas

☐ 3 a 4 horas

☐ 5 a 6 horas

☐ Mais de 6 horas

7 Como você avalia seu nível de estresse como professor?

(1 = Muito baixo | 5 = Muito alto)

8 Como você avalia sua satisfação profissional como professor?
(1 = Muito baixa | 5 = Muito alta)

9 Você já utilizou alguma ferramenta específica para criar questões anteriormente?

☐ Sim ☐ Não

10 Especificamente sobre LLMs como ChatGPT, Gemini, etc., você já teve experiência prévia em usá-los para alguma finalidade (profissional ou pessoal)??

☐ Não tive experiência prévia

☐ Sim: _____

3.5. Você acredita que o uso de LLMs pode ser uma ferramenta útil para professores na criação de exercícios?

☐ Sim

☐ Não

Seção 2: Avaliação das Questões Geradas

Nesta seção, você avaliará um conjunto de questões de Língua Portuguesa. Para cada questão, por favor, atribua uma pontuação de 1 a 5 para cada um dos cinco critérios definidos anteriormente (Cobertura, Gramaticalidade, Utilidade, Relevância, Capacidade de Resposta).

Cobertura: Avalia se a questão aborda os pontos essenciais do conteúdo esperado.

1: Poucos ou nenhum elemento relevante coberto no contexto.

2: Alguns elementos são cobertos no contexto, mas detalhes importantes estão ausentes.

3: Vários elementos importantes são cobertos no contexto, mas alguns detalhes são negligenciados.

4: Cobertura ampla dos elementos essenciais do contexto.

5: Cobertura abrangente de todos os elementos essenciais do contexto.

Gramaticalidade: Verifica se a questão está escrita corretamente, seguindo as normas gramaticais e sendo clara.

- 1: Numerosos erros; difícil de entender.
- 2: Vários erros; o significado é parcialmente obscuro.
- 3: Alguns erros, geralmente compreensíveis.
- 4: Erros menores; maior parte clara e correta.
- 5: Sem erros; completamente clara e correta.

Utilidade: Mede o valor prático da questão para avaliar ou ensinar o aluno, mesmo que não seja totalmente respondível com base no contexto.

- 1: Não útil; nenhuma contribuição para a aprendizagem.
- 2: Utilidade limitada; impacto mínimo na avaliação.
- 3: Moderadamente útil; contribui para a aprendizagem, embora seja um pouco previsível.
- 4: Útil; melhora a avaliação e a compreensão do estudante.
- 5: Altamente útil; aumenta significativamente a aprendizagem e os resultados da avaliação.

Relevância: Analisa se a questão está diretamente relacionada ao tema ou contexto proposto, sem desvios.

- 1: Não relevante; sem conexão com o contexto.
- 2: Pouco relevante; conexão fraca com o contexto.
- 3: Moderadamente relevante; algum foco no contexto.
- 4: Relevante; alinhada de perto com o contexto.
- 5: Altamente relevante; aborda diretamente o contexto sem ideias não relacionadas.

Capacidade de resposta: Determina se é possível responder à questão usando a

lógica ou as informações do contexto fornecido.

1: Não respondível; sem base no contexto.

2: Difícil de responder; conexão fraca com o contexto.

3: Moderadamente respondível; alguma lógica pode ser aplicada a partir do contexto.

4: Quase respondível; base clara no contexto.

5: Totalmente respondível; diretamente suportada pelo contexto.

Questão:

[Inserir aqui o texto completo da questão gerada a ser avaliada - SEM A FONTE]

Avalie a questão nos seguintes critérios:

Cobertura Curricular: (1: Muito Baixa - 5: Muito Alta)

Gramaticalidade: (1: Muitos Erros - 5: Sem Erros)

Utilidade Pedagógica: (1: Não Útil - 5: Altamente Útil)

Relevância Contextual: (1: Não Relevante - 5: Altamente Relevante)

Capacidade de Resposta: (1: Não Respondível - 5: Totalmente Respondível)

Comentários Adicionais sobre a Questão [Código da Questão] (Opcional):

(Repetir o bloco abaixo para CADA questão que será avaliada)

Apêndice D – E-mail para convidar os Professores

Assunto: Convite para Participação em Pesquisa sobre Geração de Questões com Inteligência Artificial (UNIRIO)

Prezado(a) Professor(a),

Meu nome é Wendy Silva Cataldo e sou mestranda no Programa de Pós-Graduação em Informática da Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Estou conduzindo uma pesquisa sob a orientação do Prof. Dr. Carlos Eduardo De Mello e coorientação da Profa. Dra. Laura de Oliveira Fernandes Moraes, intitulada "Estudo sobre o uso de Meta-prompting em LLMs para a Elaboração de Exercícios: Na Perspectiva de Docentes".

Gostaríamos de convidá-lo(a) a participar desta pesquisa respondendo a um questionário online. Sua participação envolverá duas etapas principais:

1. Responder a algumas perguntas sobre seu perfil profissional, experiência docente e percepções sobre tecnologias educacionais.
2. Avaliar um conjunto de questões de Língua Portuguesa (apresentadas anonimamente) com base em critérios específicos (Cobertura, Gramaticalidade, Utilidade, Relevância e Capacidade de Resposta).

Estimamos que o preenchimento completo do questionário levará aproximadamente 20 minutos.

Asseguramos que sua participação é voluntária e todas as suas respostas serão tratadas com estrita confidencialidade e anonimato, sendo utilizadas exclusivamente para fins acadêmicos.

Para participar, por favor, acesse o questionário através do seguinte link:

Link: <https://forms.gle/VVzjCU8NRp8U4Je76>

O questionário inclui um Termo de Consentimento Livre e Esclarecido (TCLE) detalhado na página inicial, que pedimos que leia atentamente antes de prosseguir.

Caso tenha qualquer dúvida sobre a pesquisa, sinta-se à vontade para entrar em contato conosco pelos e-mails: wendy.cataldo@edu.unirio.br.

Agradeço imensamente sua atenção e potencial colaboração.

Atenciosamente,

Wendy Silva Cataldo

Mestranda em Informática – UNIRIO

Apêndice E - Resultados Detalhados

Para examinar a relação entre nível de estresse e as avaliações, consulte a tabela E.1.

Tabela E.1 - Critérios por faixa de tempo de docência (mediana, Q1, Q3, IQR e n).

Critério	Tempo de docência	n	Mediana	Q1	Q3	IQR
Capacidade de resposta	<1	5	4,00	4,00	4,00	0,00
Capacidade de resposta	1–5	4	3,50	2,75	4,25	1,50
Capacidade de resposta	6–10	16	5,00	5,00	5,00	0,00
Capacidade de resposta	11–15	23	4,50	3,75	5,00	1,25
Capacidade de resposta	16–20	13	5,00	4,00	5,00	1,00
Capacidade de resposta	>20	20	5,00	4,00	5,00	1,00
Cobertura	<1	5	4,00	3,50	4,50	1,00
Cobertura	1–5	4	4,00	4,00	4,00	0,00
Cobertura	6–10	16	5,00	4,00	5,00	1,00
Cobertura	11–15	23	4,00	4,00	5,00	1,00
Cobertura	16–20	13	4,00	3,00	5,00	2,00
Cobertura	>20	20	4,25	3,88	5,00	1,12
Gramaticalidade	<1	5	5,00	5,00	5,00	0,00
Gramaticalidade	1–5	4	3,50	3,00	4,25	1,25
Gramaticalidade	6–10	16	5,00	5,00	5,00	0,00
Gramaticalidade	11–15	23	5,00	3,25	5,00	1,75
Gramaticalidade	16–20	13	4,00	3,00	5,00	2,00
Gramaticalidade	>20	20	5,00	3,75	5,00	1,25

Relevância	<1	5	4,00	4,00	4,50	0,50
Relevância	1–5	4	3,50	2,75	4,12	1,38
Relevância	6–10	16	5,00	5,00	5,00	0,00
Relevância	11–15	23	4,50	3,75	5,00	1,25
Relevância	16–20	13	4,00	3,00	5,00	2,00
Relevância	>20	20	5,00	3,88	5,00	1,12
Utilidade	<1	5	4,00	3,50	4,00	0,50
Utilidade	1–5	4	3,50	3,00	4,25	1,25
Utilidade	6–10	16	5,00	5,00	5,00	0,00
Utilidade	11–15	23	4,00	4,00	5,00	1,00
Utilidade	16–20	13	3,50	3,00	5,00	2,00
Utilidade	>20	20	4,25	3,88	5,00	1,12

Nota: Escala 1–5. Mediana = valor central; Q1 = 1º quartil; Q3 = 3º quartil; IQR = Q3 – Q1 (dispersão; menor = maior consistência).

Fonte: Elaborada pela autora (2025).

A tabela E.1 consolida, por tempo de docência (<1, 1–5, 6–10, 11–15, 16–20, >20 anos), as notas por professor em cinco critérios (Capacidade de resposta, Cobertura, Gramaticalidade, Relevância, Utilidade), apresentando n (tamanho do grupo), Mediana (nível central), Q1 e Q3 (quartis que delimitam a faixa central de 50% das notas) e IQR (Q3–Q1, medida de dispersão). Em Capacidade de resposta, as medianas vão de 3,50 (1–5 anos; IQR=1,50) a 5,00 (6–10, 16–20, >20 anos; IQRs 0,00–1,00), com 4,50 em 11–15 (IQR=1,25) e 4,00 em <1 (IQR=0,00). Em Cobertura, a mediana é 5,00 em 6–10 (IQR=1,00); 4,25 em >20 (IQR=1,12); e 4,00 nas demais faixas (com IQR de 0,00 a 2,00 – destaque para 1–5 com IQR=0,00 e 16–20 com IQR=2,00). Em Gramaticalidade, a mediana é 5,00 em <1 e 6–10 (IQR=0,00 em ambas), 11–15 (IQR=1,75) e >20 (IQR=1,25); 4,00 em 16–20 (IQR=2,00); e 3,50 em 1–5 (IQR=1,25). Em Relevância, a mediana atinge 5,00 em 6–10 (IQR=0,00) e >20 (IQR=1,12), é 4,50 em 11–15 (IQR=1,25), 4,00 em <1 e 16–20 (IQR=0,50 e 2,00,

respectivamente) e 3,50 em 1–5 (IQR=1,38). Em Utilidade, registra-se 5,00 em 6–10 (IQR=0,00), 4,25 em >20 (IQR=1,12), 4,00 em <1 e 11–15 (IQR=0,50 e 1,00) e 3,50 em 1–5 e 16–20 (IQR=1,25 e 2,00). Em praticamente todos os critérios e faixas, Q3 alcança 5,00 (efeito teto), com exceções pontuais (ex.: <1 ano em Cobertura e Utilidade).

O padrão não mostra linha crescente/decrescente uniforme com a senioridade. Observa-se um pico de desempenho e aceitação em 6–10 anos (medianas 5,00 e IQR=0,00 em vários critérios), seguido por um alto patamar em >20 anos (medianas altas com dispersão moderada). A faixa 11–15 anos mantém nível central elevado, mas com variação entre avaliadores (IQR maior). Em contraste, 1–5 anos agrega notas mais baixas e dispersas, o que pode refletir diferenças de experiência e critérios em estágios iniciais (agravado por n pequeno). Já 16–20 anos apresenta a maior diferença (IQR=2,00 em quatro critérios), sugerindo contextos e expectativas mais diversas nessa fase. As diferenças entre faixas existem, mas o melhor desempenho e a maior consistência tendem a aparecer no meio de carreira (6–10 anos), com efeito teto frequente (Q3=5,00) e ausência de relação linear entre tempo de docência e as notas.

Para a hipótese de que quem tem Mestrado/Doutorado tende a dar notas mais altas em Gramaticalidade e Relevância e a ser mais consistente, podemos ver a tabela E.2 e E.3.

Tabela E.2 - Gramaticalidade e Relevância por titulação (mediana, Q1, Q3, IQR e n).

Critério	Titulação	n	Mediana	Q1	Q3	IQR
Gramaticalidade	Doutorado	14	5,00	3,12	5,00	1,88
Gramaticalidade	Especialização	34	5,00	4,50	5,00	0,50
Gramaticalidade	Mestrado	22	5,00	4,12	5,00	0,88
Gramaticalidade	Sem pós	11	4,00	3,00	4,75	1,75
Relevância	Doutorado	14	4,25	3,12	5,00	1,88
Relevância	Especialização	34	5,00	4,00	5,00	1,00

Relevância	Mestrado	22	4,75	4,00	5,00	1,00
Relevância	Sem pós	11	4,00	3,00	4,75	1,75

Nota: Escala 1–5. Mediana = valor central; Q1 = 1º quartil; Q3 = 3º quartil; IQR = Q3 – Q1 (dispersão; menor = maior consistência).

Fonte: Elaborada pela autora (2025).

A tabela E.2 reúne, para Gramaticalidade e Relevância, as estatísticas das notas por professor, agrupadas por titulação: n (número de professores), Mediana (nível central), Q1 e Q3 (quartis que delimitam a faixa central de 50% das notas) e IQR (amplitude interquartil, Q3–Q1). Em Gramaticalidade, a mediana = 5,00 em Doutorado (n=14; IQR=1,88), Especialização (n=34; IQR=0,50) e Mestrado (n=22; IQR=0,88); em Sem pós (n=11), a mediana é 4,00 (IQR=1,75). Em Relevância, as medianas são 5,00 em Especialização (n=34; IQR=1,00), 4,75 em Mestrado (n=22; IQR=1,00), 4,25 em Doutorado (n=14; IQR=1,88) e 4,00 em Sem pós (n=11; IQR=1,75). Em ambos os critérios, os *quartis* superiores (Q3) alcançam 5,00 em todas as titulações, indicando teto elevado de avaliação.

As medianas são altas em todos os níveis com pós-graduação, especialmente em Especialização e Mestrado. Em Gramaticalidade, há empate no teto (5,00) entre Doutorado, Mestrado e Especialização, com dispersão menor em Especialização (IQR=0,50) e Mestrado (IQR=0,88) do que em Doutorado (IQR=1,88). Em Relevância, o pico está em Especialização (5,00), seguido de Mestrado (4,75); Doutorado (4,25) fica abaixo, sugerindo que maior titulação não implica necessariamente maior nota para esse critério. Sem pós apresenta menores medianas e maior dispersão em ambos os critérios. No conjunto, há diferenças entre titulações, mas sem tendência unidirecional de “maior titulação levando a notas mais elevadas”; o desempenho “alto e estável” se concentra em Especialização/Mestrado, enquanto Doutorado mostra medianas altas, porém mais heterogeneidade.

Tabela E.3.IQR - Consistência (IQR do avaliador) por titulação e critério (mediana_IQR, Q1_IQR, Q3_IQR e n).

Critério	Titulação	n	Mediana_IQR	Q1_IQR	Q3_IQR
Gramaticalidade	Sem pós	11	1,00	0,50	1,62
Gramaticalidade	Especialização	34	0,00	0,00	0,25
Gramaticalidade	Mestrado	22	0,25	0,00	1,00
Gramaticalidade	Doutorado	14	1,00	0,06	1,00
Relevância	Sem pós	11	1,00	1,00	1,25
Relevância	Especialização	34	0,25	0,00	1,00
Relevância	Mestrado	22	1,00	0,25	1,25
Relevância	Doutorado	14	1,00	0,44	1,19

Nota: IQR do avaliador mede a dispersão das notas dadas por cada professor (menor = maior consistência). Estatísticas apresentadas: mediana_IQR, Q1_IQR, Q3_IQR.

Fonte: Elaborada pela autora (2025).

A tabela E.3 resume, para Gramaticalidade e Relevância, a consistência das avaliações por titulação, medida pelo IQR do avaliador: mediana_IQR (nível central da dispersão), Q1_IQR e Q3_IQR (faixa central de 50% das dispersões) e n (tamanho do grupo). Em Gramaticalidade, Especialização apresenta a menor dispersão (mediana_IQR=0,00; Q3_IQR=0,25), seguida de Mestrado (0,25; 1,00); Sem pós (1,00; 1,62) e Doutorado (1,00; 1,00) exibem maior dispersão. Em Relevância, Especialização novamente mostra menor dispersão (0,25; 1,00), enquanto Mestrado, Doutorado e Sem pós ficam com mediana_IQR=1,00 (Q3_IQR entre 1,19 e 1,25).

A consistência (menor IQR) tende a ser melhor em Especialização (e, em menor grau, em Mestrado), sobretudo em Gramaticalidade. Em Relevância, o padrão se repete: Especialização mantém dispersão menor, enquanto demais níveis apresentam IQR mediano=1,00. Isso indica que, embora as medianas de desempenho sejam altas em vários níveis, a estabilidade das avaliações é superior entre docentes com Especialização (e razoável em Mestrado), ao passo que

Doutorado e Sem pós mostram maior variabilidade. Em conjunto com a tabela E.3, conclui-se que não há ganho unidirecional “titulações mais altas associam-se a medianas superiores e menor IQR”; em vez disso, Especialização/Mestrado combinam alto nível central com maior consistência, enquanto Doutorado mantém desempenho alto, porém mais heterogêneo, e Sem pós fica abaixo em mediana e acima em dispersão.

Para a hipótese de que se esperam diferenças entre escolas pública, privada e militar nas notas de Utilidade e Cobertura, podemos ver a tabela E.4.

Tabela E.4 - Utilidade e Cobertura por rede (mediana, Q1, Q3, IQR e n).

Critério	Rede	n	Mediana	Q1	Q3	IQR
Utilidade	Militar	5	3,00	3,00	4,50	1,50
Utilidade	Privada	12	4,75	4,00	5,00	1,00
Utilidade	Pública	42	4,00	3,62	5,00	1,38
Cobertura	Militar	5	4,00	3,50	4,50	1,00
Cobertura	Privada	12	4,00	3,88	5,00	1,12
Cobertura	Pública	42	4,00	4,00	5,00	1,00

Nota: Escala 1–5. Mediana = valor central; Q1 = 1º quartil; Q3 = 3º quartil; IQR = Q3 – Q1 (dispersão; menor = maior consistência).

Fonte: Elaborada pela autora (2025).

A tabela E.4 apresenta, para Utilidade e Cobertura, as estatísticas das notas por professor, agrupadas por rede (Militar, Privada, Pública): n (tamanho do grupo), Mediana (nível central), Q1 e Q3 (quartis da faixa central de 50%) e IQR (amplitude interquartil, Q3–Q1). Em Utilidade, as medianas são 4,75 (Privada; $n=12$; IQR=1,00), 4,00 (Pública; $n=42$; IQR=1,38) e 3,00 (Militar; $n=5$; IQR=1,50), com Q3=5,00 em Privada e Pública e Q3=4,50 em Militar. Em Cobertura, as três redes apresentam mediana = 4,00, com IQR próximo de 1,00 (Militar=1,00; Privada=1,12; Pública=1,00) e Q3=5,00 para Privada e Pública (Militar: Q3=4,50).

O padrão por rede difere entre os critérios. Em Utilidade, há vantagem aparente

da rede Privada (mediana 4,75 e menor dispersão relativa, IQR=1,00), seguida da rede Pública (mediana 4,00, IQR=1,38) e, por fim, da rede Militar (mediana 3,00, IQR=1,50). Ainda assim, há sobreposição de *quartis* (Pública e Privada com Q3=5,00), e o n muito menor da rede Militar recomenda cautela na comparação. Em Cobertura, as três redes exibem mediana idêntica (4,00) e dispersões muito próximas (IQR≈1,00), com Q3=5,00 em Privada e Pública, sugerindo avaliações semelhantes entre redes para a abrangência do conteúdo. No conjunto, os resultados não sustentam diferenças sistemáticas por rede em Cobertura e indicam apenas uma vantagem pontual da rede Privada em Utilidade. Considerando a sobreposição de quartis e os tamanhos amostrais desequilibrados, a hipótese de diferenças por rede não se confirma no agregado, embora Privada tenda a avaliar a Utilidade de forma mais alta e ligeiramente mais consistente.

Para a hipótese de que quem elabora questões com mais frequência e dedica mais horas tende a ser mais crítico em Utilidade, podemos ver na tabela E.5.

Tabela E.5 - Utilidade para prática de elaboração (frequência e horas semanais): mediana, Q1, Q3, IQR e n.

Dimensão	Nível (0–4)	n	Mediana	Q1	Q3	IQR
Frequência de elaboração	0	7	3,00	3,00	3,50	0,50
Frequência de elaboração	1	5	4,00	4,00	5,00	1,00
Frequência de elaboração	2	14	4,00	3,00	4,00	1,00
Frequência de elaboração	3	25	4,50	4,00	5,00	1,00
Frequência de elaboração	4	30	5,00	4,00	5,00	1,00
Horas semanais	0	5	3,50	3,00	3,50	0,50
Horas semanais	1	21	4,00	4,00	5,00	1,00
Horas semanais	2	30	4,00	3,50	5,00	1,50
Horas semanais	3	15	5,00	3,00	5,00	2,00
Horas semanais	4	10	5,00	4,25	5,00	0,75

Nota: Escala 1–5. Mediana = valor central; Q1 = 1º quartil; Q3 = 3º quartil; IQR = Q3 – Q1

(dispersão; menor = maior consistência). Codificação dos níveis: Frequência (0 = Nunca ... 4 = Sempre); Horas/semana (0 = <1 ... 4 = >6).
 Fonte: Elaborada pela autora (2025).

A tabela E.5 apresenta, para o critério Utilidade, as estatísticas das notas por professor agrupadas por frequência de elaboração de questões e por horas semanais dedicadas: *n* (tamanho do grupo), Mediana (nível central), Q1 e Q3 (quartis que delimitam a faixa central de 50% das notas) e IQR (amplitude interquartil, Q3–Q1). Em Frequência, as medianas crescem de 3,00 no nível 0 (*n*=7; IQR=0,50) para 4,00 nos níveis 1 (*n*=5; IQR=1,00) e 2 (*n*=14; IQR=1,00), alcançando 4,50 no nível 3 (*n*=25; IQR=1,00) e 5,00 no nível 4 (*n*=30; IQR=1,00). Em Horas semanais, as medianas evoluem de 3,50 no nível 0 (*n*=5; IQR=0,50) para 4,00 nos níveis 1 (*n*=21; IQR=1,00) e 2 (*n*=30; IQR=1,50), atingindo 5,00 nos níveis 3 (*n*=15; IQR=2,00) e 4 (*n*=10; IQR=0,75). Em ambos os eixos, Q3 frequentemente chega a 5,00, indicando teto elevado nas avaliações superiores.

O padrão é claro e não unidirecional apenas na dispersão: a mediana aumenta conforme cresce a frequência de elaboração (3,00 para 5,00 do nível 0 ao 4) e também com mais horas semanais (3,50 para 5,00 do nível 0 ao 3/4). Isso contraria a hipótese de que mais prática/tempo implicaria maior exigência e, portanto, notas menores em Utilidade. Na frequência, a consistência é relativamente estável (IQR $\approx$ 1,00) dos níveis 1 a 4, com concordância alta no nível 0 (IQR=0,50) embora com *n* pequeno, sugerindo cautela. Nas horas semanais, há maior diferença no nível 3 (IQR=2,00) e consistência retomada no nível 4 (IQR=0,75), o que pode refletir perfis distintos (ex.: docentes que investem muito tempo por necessidade de adequação vs. por dominar o processo). A maior prática medida por frequência e por tempo dedicado está associada a avaliações mais altas de Utilidade, com dispersão moderada na maior parte dos grupos. A hipótese de “mais prática/mais tempo dedicados estariam associados à maior rigor (ou seja, notas mais baixas)” é, portanto, refutada neste conjunto.

Para a hipótese de que quem acredita que LLMs são úteis tende a dar notas mais altas às questões geradas por LLM, mesmo sem experiência prévia, podemos ver a tabela E.6.

Tabela E.6 — Mediana global por crença na utilidade dos LLMs (geral e estrato “sem experiência”) — mediana, Q1, Q3, IQR e n.

Estrato	Crença	n	Mediana	Q1	Q3	IQR
Geral	Não	5	4,00	3,00	4,00	1,00
Geral	Sim	76	5,00	4,00	5,00	1,00
Sem experiência	Não	5	4,00	3,00	4,00	1,00
Sem experiência	Sim	46	5,00	4,00	5,00	1,00

Nota: Escala 1–5. Mediana = valor central; Q1 = 1º quartil; Q3 = 3º quartil; IQR = Q3 – Q1 (dispersão; menor = maior consistência).

Fonte: Elaborada pela autora (2025).

A tabela E.6 apresenta a mediana global das avaliações (escala 1–5) segundo a crença na utilidade dos LLMs, no conjunto geral e no estrato “sem experiência prévia”. Em ambos os recortes, quem crê na utilidade (Sim) exhibe mediana = 5,00 ($Q1=4,00$; $Q3=5,00$; $IQR=1,00$), enquanto o grupo Não fica em 4,00 ($Q1=3,00$; $Q3=4,00$; $IQR=1,00$). Observa-se teto ($Q3=5,00$) no grupo Sim e IQR idêntico (1,00) entre as crenças, com amostras muito desiguais ($n=76/46$ para *Sim* vs. $n=5$ para *Não*).

O padrão é consistente na crença na utilidade dos LLMs associa-se a notas globais mais altas (mediana 5,00 vs. 4,00), e isso permanece quando se restringe a análise aos docentes sem experiência prévia. Como os IQRs são iguais (1,00), a diferença está no nível central e não na dispersão, quem crê tende a posicionar as avaliações um ponto acima na escala, com efeito teto frequente ($Q3=5,00$). Duas cautelas: (i) o grupo Não é muito pequeno ($n=5$), o que limita a precisão de comparações; (ii) a presença de teto pode atenuar diferenças adicionais entre grupos. Os dados sustentam a hipótese, inclusive entre quem ainda não tem experiência com LLMs.

Para a hipótese de que mais estresse tende a vir com notas mais baixas ou maior variação nas avaliações, podemos ver a tabela E.7.

Tabela E.7 — Utilidade e Relevância por grupo de estresse (mediana, Q1, Q3, IQR e n).

Critério	Grupo de estresse	n	Mediana	Q1	Q3	IQR
Utilidade	Baixo(1–2)	13	5,00	4,00	5,00	1,00
Utilidade	Médio(3)	21	5,00	4,00	5,00	1,00
Utilidade	Alto(4–5)	47	4,00	3,50	5,00	1,50
Relevância	Baixo(1–2)	13	5,00	4,00	5,00	1,00
Relevância	Médio(3)	21	5,00	4,00	5,00	1,00
Relevância	Alto(4–5)	47	4,00	3,50	5,00	1,50

Nota: Escala 1–5. Mediana = valor central; Q1 = 1º quartil; Q3 = 3º quartil; IQR = Q3 – Q1 (dispersão; menor = maior consistência). Codificação: Baixo(1–2), Médio(3), Alto(4–5).
Fonte: Elaborada pela autora (2025).

A tabela E.7 apresenta, para Utilidade e Relevância, as estatísticas das notas por professor, agrupadas por nível de estresse: n (tamanho do grupo), Mediana (nível central), Q1 e Q3 (quartis que delimitam a faixa central de 50% das notas) e IQR (amplitude interquartil, Q3–Q1). Em Baixo(1–2) e Médio(3), a mediana = 5,00 com IQR=1,00 em ambos os critérios; no grupo Alto(4–5), a mediana = 4,00, Q1=3,50, Q3=5,00 e IQR=1,50. Em todas as combinações, Q3 alcança 5,00, indicando presença de teto nas avaliações superiores.

O padrão é consistente com a hipótese maior estresse associa-se a notas mais baixas (mediana 4,00 no grupo Alto) e a maior dispersão (IQR=1,50), sugerindo diferença de julgamento sob maior pressão. Já os grupos Baixo e Médio exibem nível central mais alto (5,00) e dispersão menor (IQR=1,00). A presença de efeito teto (Q3=5,00) em todos os grupos indica que uma parcela dos avaliadores atribui notas máximas, o que pode atenuar diferenças adicionais. Além disso, os tamanhos amostrais são desbalanceados (ex: $n=47$ no grupo Alto), o que requer cautela na comparação direta. Os dados confirmam que o estresse mais alto está ligado a notas mais baixas e maior variabilidade nas avaliações de Utilidade e Relevância.

Apêndice F – Manual para os professores

Este guia oferece orientações operacionais para criar e revisar exercícios com LLMs usando Prompt e Meta-Prompt. À luz dos resultados desta dissertação. Por isso, as recomendações abaixo priorizam especificidade do pedido, contexto pedagógico e curadoria docente.

Escopo e premissas

- Público-alvo: docentes de Língua Portuguesa e áreas afins.
- Tarefa: geração e revisão de questões alinhadas à BNCC e a objetivos de aprendizagem claros.
- Princípios do guia: Clareza do objetivo e contexto superam camadas meta genéricas. Curadoria docente é indispensável (adequação curricular, linguagem, viés, dificuldade).

Quando usar cada abordagem

- Prompt: Use quando você já sabe o que pedir (conteúdo, série, nível de Bloom, formato de saída). Na maioria dos casos, é suficiente para gerar itens adequados.
- Meta-Prompt: Use quando você ainda não tem um enunciado claro e quer que o LLM ajude a formular o melhor prompt, antes de gerar os itens. Trate como ferramenta de variação, não como garantia de melhora.

Nota baseada nos achados: nas comparações do estudo, Prompt vs. Meta-Prompt apresentaram desempenho comparável. A escolha deve considerar tempo e clareza que você já tem sobre o objetivo didático.

Fluxo de trabalho recomendado

1. Defina o objetivo de aprendizagem (BNCC + nível de Bloom) e o conteúdo (ex.: coesão referencial, concordância, interpretação).

2. Escolha a abordagem: prompt ou meta-prompt (se precisar de ajuda para estruturar o pedido).
3. Gere o rascunho de itens com gabarito comentado e graduação de dificuldade (fácil/médio/difícil).
4. Aplique a curadoria docente (checklist).
5. Ajuste e finalize (linguagem inclusiva, adequação etária, referências/trechos de apoio quando houver).

Checklist de curadoria docente

- Alinhamento curricular: incluir código BNCC pertinente e nível de Bloom.
- Adequação linguística: gramática/ortografia; adequação etária e cultural; linguagem inclusiva.
- Relevância e cobertura: o item avalia exatamente o que se propõe? Cobre o conteúdo indicado?
- Gabarito e justificativa: a resposta correta é inequívoca? A justificativa é clara e fiel ao conteúdo?
- Dificuldade e progressão: há itens fácil/médio/difícil? A escala de complexidade é coerente?
- Viés e sensibilidade: evitar vieses socioeconômicas, culturais, regionais ou de gênero; revisar estereótipos.
- Formato e clareza: enunciado objetivo, comandos explícitos, tempo estimado realista.
- Direitos autorais e fontes: se usar trechos de textos, cite fontes e verifique direitos.

Dica: mantenha este checklist como rubrica para revisão rápida de qualquer

item, seja gerado por LLM ou humano.

Modelos prontos de prompts

Prompt

Objetivo: Gerar 3 questões de Língua Portuguesa para [ano/série] sobre [conteúdo], alinhadas à BNCC [código] e ao nível [Bloom].

Requisitos: Inclua gabarito comentado, graduação de dificuldade (fácil/médio/difícil) e indique tempo estimado por item. Use linguagem inclusiva e evite vieses. Se usar texto-base, cite a fonte e insira o trecho no enunciado.

Formato de saída:

1. Enunciado
2. Alternativas (se múltipla escolha) ou resposta esperada (se discursiva)
3. Gabarito comentado (1–3 linhas)
4. Dificuldade (F/M/D) e tempo estimado

Meta-prompt

Etapa A — Construir o melhor prompt

Atue como professor(a) de Língua Portuguesa. Proponha o melhor prompt para gerar 3 questões sobre [conteúdo] para [ano/série], alinhadas à BNCC [código] e nível [Bloom]. O prompt final deve exigir gabarito comentado, graduação de dificuldade (F/M/D) e tempo estimado. Inclua lembretes de linguagem inclusiva, evitar vieses e citar fonte quando houver texto-base.

- “Inclua pelo menos um item de análise linguística é um item interpretativo.”
- “Inclua um item contextualizado em [gênero textual real] (ex.: notícia, crônica, HQ curta).”

- “Inclua um item com erro proposital para identificação/correção pelo aluno.”

Etapas B — Gerar as questões com o prompt aprovado

Agora use o prompt final e gere as 3 questões. Siga o formato de saída da Seção E.5.1.

Variações úteis:

- “Gere duas versões alternativas do prompt final com abordagens distintas (ex.: foco em coesão vs. foco em pontuação), depois avalie qual é mais adequada ao objetivo [descrever objetivo].”
- “Gere variações de enunciado para o mesmo foco BNCC/Bloom, mantendo o gabarito.”

Boas práticas e limites

- Evite depender de meta-prompts genéricos esperando “ganhos automáticos”.
- Priorize: objetivo claro, contexto curricular (BNCC) e exemplos.
- Revise sempre com o checklist especialmente gabarito e viés.
- Teste em pequena escala (1 turma, 1 atividade) antes de adoção ampla.

Ética, vieses e linguagem inclusiva

- Evitar vieses: socioeconômicos, culturais, regionais, de gênero/raça; revisar estereótipos ou generalizações.
- Citar fontes quando usar textos base; respeitar direitos autorais.
- Não usar dados pessoais de estudantes que permitam identificação.
- Adequação etária: evitar temas sensíveis sem mediação didática e autorização escolar.