Uma análise de qualidade do uso de grandes modelos de linguagem para geração automática de itens avaliativos em português

João Vítor de Castro Martins Ferreira Nogueira[1,3], João Augusto Pilato de Castro[1,3], Lucas O. Larcher[1,3], Rosângela Veiga Júlio Ferreira[2,3], Begma Tavares Barbosa[2,3], Jairo Francisco de Souza[1,2,3]

[1]LApIC Research Group – UFJF – MG — Brasil

[2]Universidade Federal de Juiz de Fora (UFJF) -– MG — Brasil

[3]Fundação CAEd – Juiz de Fora – MG – Brasil

{pilato.joao, joao.nogueira}@estudante.ufjf.br,
{jairo.souza, rosangelaveiga.ferreira}@ufjf.br,
begmatb@gmail.com, lucas.larcher@fundacaocaed.org.br

Abstract. *Educational assessment plays a fundamental role in monitoring the quality of education. The manual creation and development of test items is a costly and highly specialized task, and the use of large language models (LLMs) has become a popular solution in the field of automatic item generation (AIG). This study proposes a quantitative and qualitative analysis of the use of general-purpose LLMs for generating Portuguese language items intended for large-scale Brazilian assessments, which have well-defined requirements for item construction and quality verification. The results indicate that current technologies are not yet capable of fully addressing this problem and that significant research challenges remain in this area.*

Resumo. *A avaliação educacional cumpre um papel fundamental no acompanhamento da qualidade da educação. A criação e elaboração manual de itens de prova é uma tarefa custosa e altamente especializada, e o uso de grandes modelos de linguagem (LLM) tem se tornado uma solução popular na área de geração automática de itens (AIG). Este estudo realiza uma analisa qualitativa e quantitativa do uso das LLMs de propósito geral na geração de itens em Língua Portuguesa para uso em avaliações brasileiras em larga escala, os quais possuem requisitos bem definidos para sua construção e verificação de qualidade. Os resultados demonstram que tecnologias atuais não são capazes de resolver completamente o problema e que há desafios de pesquisa nesta área.*

1. Introdução

A avaliação educacional cumpre um papel essencial como instrumento de diagnóstico do processo de aprendizagem, permitindo que o direito à educação de qualidade seja monitorado, analisado e continuamente aprimorado [Silva et al. 2016]. Para assegurar a qualidade do processo de aprendizagem, é fundamental dispor de mecanismos confiáveis de mensuração, como os testes. No contexto das avaliações em larga escala — aquelas aplicadas a um grande número de estudantes em âmbitos municipal, estadual ou nacional

DOI: 10.5753/sbie.2025.12210

[Greaney and Kellaghan 2008] — a elaboração desses instrumentos constitui uma tarefa complexa, que envolve desde a definição das habilidades a serem avaliadas e das formas de aferição, até a construção e diversificação de seus componentes centrais: os itens de prova, que devem apresentar diferentes níveis de dificuldade e adequação pedagógica.

O teste é um instrumento de avaliação que descreve, numericamente, o grau de aprendizado em condições padronizadas. Um item, unidade básica do teste, fornece uma instrução ao estudante, cuja resposta pode ser pontuada de diferentes formas [Haladyna 2004]. Em avaliações em larga escala, cada item mede uma habilidade específica da matriz de referência e exige alta especialização: deve ser pedagógica e linguisticamente adequado, tecnicamente válido e sensível ao desempenho estudantil. A elaboração manual demanda tempo, profissionais qualificados e ciclos de revisão, tornando-se um gargalo logístico diante da crescente demanda por avaliações frequentes, diversificadas e alinhadas a diferentes contextos curriculares [Gierl and Lai 2018]. Nesse cenário, a Geração Automática de Itens (Automatic Item Generation – AIG) surge como solução promissora, combinando geração computacional de texto com curadoria pedagógica para agilizar a produção, ampliar a diversidade de formatos e possibilitar personalização de avaliações [Gierl et al. 2021, Circi et al. 2023].

No entanto, apesar do potencial que os modelos de linguagem apresentam para a AIG ao aprimorar a geração de textos, as abordagens predominantes para avaliar a qualidade dos itens gerados se concentram em métricas automáticas, baseadas em similaridade textual ou semântica, que permitem a avaliação de grandes volumes de itens [Song et al. 2025]. Contudo, a qualidade de um item está ligada a outros pontos, como validade pedagógica, clareza, adequação ao nível de escolaridade e potencial diagnóstico. Tais atributos dificilmente podem ser aferidos apenas por métricas automáticas [Maity et al. 2024, Scaria et al. 2024]. Nesse sentido, a avaliação humana especializada continua sendo insubstituível para validar a efetividade dos itens [Setiawan et al. 2022]. Embora alguns trabalhos, como os de [Lin and Chen 2024, Scaria et al. 2024] tenham recorrido a especialistas para avaliar os itens gerados, e outros, como [Rockembach and Thom 2024, Zimerman et al. 2024], tenham utilizado avaliação humana no contexto da AIG em português brasileiro, ainda não existem critérios consolidados que guiem a avaliação da qualidade dos itens em contextos de larga escala no Brasil. Além disso, não foram encontrados trabalhos que avaliam os resultados por meio de especialistas em língua portuguesa ou da área de educação, segundo as diretrizes nacionais. Também não foram encontrados trabalhos que estudam a AIG no contexto de avaliação em larga escala em língua portuguesa.

Este trabalho tem como objetivo analisar a qualidade de itens gerados por abordagens de engenharia de *prompt* em grandes modelos de linguagem (Large Language Models – LLM), voltados à avaliação de Língua Portuguesa, a partir da perspectiva de especialistas em avaliação educacional. Para isso, é proposta uma análise qualitativa conduzida por profissionais com experiência na elaboração de itens para avaliações brasileiras em larga escala. A análise leva em consideração um conjunto de critérios baseados nas diretrizes utilizadas em avaliações brasileiras. As principais contribuições deste estudo são: (i) apresentar uma avaliação feita por especialistas em avaliações em larga escala de língua portuguesa, respeitando as diretrizes nacionais de avaliação educacional, sobre os resultados alcançados no problema de AIG utilizando diferentes técnicas de engenharia de

prompt em grandes modelos de linguagem; e (ii) oferecer um conjunto de análises que podem orientar futuras pesquisas e práticas em AIG para o contexto da Língua Portuguesa, especialmente em avaliações em larga escala.

2. Construção de itens para avaliação de larga escala

Para que um item cumpra sua função de mensurar com precisão o domínio de uma habilidade, ele deve estar alinhado com os objetivos de aprendizagem, apresentar estrutura clara e equilibrada, e ser capaz de revelar, por meio das alternativas, os caminhos cognitivos percorridos pelos estudantes [Bollela et al. 2018]. Esse aspecto diagnóstico torna o item não apenas um instrumento de medição de desempenho, mas também um recurso essencial para identificar as dificuldades dos estudantes e orientar intervenções pedagógicas mais eficazes no processo de aprendizagem. Formalmente, existem dois tipos de itens: itens de resposta construída, que são popularmente conhecidos como questões descritivas, ou seja, que têm espaço para construir uma solução, e os itens de resposta selecionada, que são as questões de múltipla escolha. No contexto de avaliação em larga escala, os itens de resposta construída possuem uma avaliação mais demorada devido ao critério de correção ser mais complexo e depender do desenvolvimento da resposta feita pelo estudante [Palacios and de Oliveira 2022]. Por outro lado, o item de resposta selecionada é de fácil correção, limitando-se à verificação da alternativa selecionada.

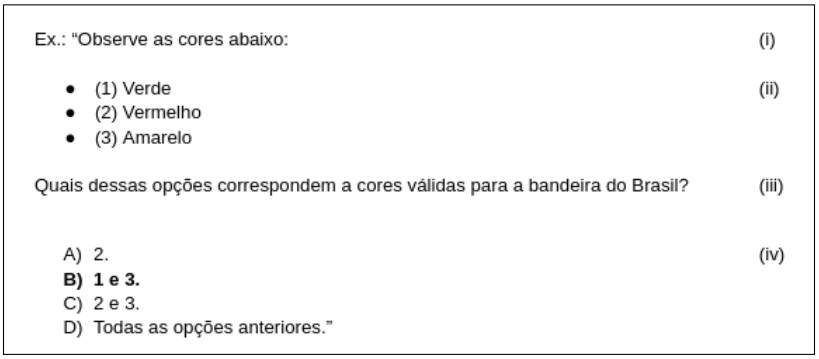

Figura 1. Exemplo de item de resposta selecionada

A Figura 1 exemplifica os principais elementos que compõem um item de resposta selecionada. O item possui comando inicial (i) que é a instrução indicando a ação que deve ser tomada para iniciar a atividade. O suporte (ii) que pode ser um texto, imagem ou outro recurso que contextualiza a questão e auxilie diretamente no desenvolvimento da resposta. O segundo comando (iii) é a instrução que indica o que o estudante deve atender para solucionar o item. As respostas (iv) consistem na união entre o gabarito (em negrito), que é a resposta correta, e os distratores, que são as respostas incorretas, mas plausíveis. A plausibilidade dos distratores é um dos principais desafios na geração de itens de múltipla escolha [Awalurahman and Budi 2024].

Dentre essas diretrizes, destacam-se quatro pontos que contribuem para a validade e qualidade dos itens: (i) conteúdo, assegurando alinhamento com os objetivos de aprendizagem e relevância pedagógica; (ii) formato do item, recomendando vocabulário simples para que a compreensão leitora não interfira na avaliação; (iii) comando claro e objetivo, com a ideia principal e instruções diretas, evitando informações irrelevantes;

(iv) alternativas plausíveis, garantindo distratores semelhantes e atrativos, evitando pistas óbvias ou que induzam à resposta correta por eliminação.

O desenvolvimento de itens válidos que atendam às necessidades pedagógicas para diagnosticar a aprendizagem dos alunos é uma tarefa complexa. Nesse contexto, em [Haladyna 2004] é apresentado diretrizes que guiam a elaboração e serão consideradas na avaliação dos itens gerados. Dentre elas, destacam-se quatro pontos-chave que contribuem para a validade e qualidade dos itens: (i) conteúdo, assegurando alinhamento com os objetivos de aprendizagem e relevância pedagógica; (ii) formato do item, recomendando vocabulário simples para evitar que a compreensão leitora interfira na avaliação; (iii) comando claro e objetivo, contendo a ideia principal e instruções diretas, sem informações irrelevantes; (iv) alternativas plausíveis, garantindo distratores semelhantes e atrativos, evitando pistas óbvias ou que induzam respostas corretas por eliminação.

A importância de seguir tais orientações é reforçada por evidências que mostram impactos negativos na elaboração de itens sem o amparo de diretrizes. Um exemplo é mostrado por [Richichi 1996], que testou um conjunto de itens do contexto de psicologia introdutória utilizando a teoria da resposta ao item (TRI) e concluiu que os itens que violaram algumas diretrizes de elaboração eram mais difíceis e menos discriminativos do que os itens que não possuíam falhas.

3. Trabalhos relacionados

Muitos trabalhos propuseram soluções de AIG envolvendo uso de LLMs com engenharia de *prompt* [Rockembach and Thom 2024, Zimerman et al. 2024, Scaria et al. 2024, Lee et al. 2024, Wang et al. 2024]. Em [Scaria et al. 2024] foram analisadas as técnicas de *prompt* que geram melhores itens, além de analisar o impacto que a complexidade e o tamanho do *prompt* na qualidade do item. A avaliação realizada por especialistas e os resultados apontaram que *prompts* mais simples tiveram piores resultados que *prompts* mais bem estruturados. Em [Rockembach and Thom 2024] foram utilizados *prompts* com e sem um template no GPT-3.5 e LLAMA-2.0 para gerar questões para uma disciplina de *Business Process Management* em português brasileiro. A avaliação foi feita pelos próprios autores e os resultados do trabalho apontam que o uso de templates melhorou a performance do LLAMA-2.0. Ressalta-se, contudo, que os critérios de qualidade para itens para uso em larga escala se diferem de itens construídos para atividades em sala de aula. Em [Bezirhan and von Davier 2023] foi proposto um modelo para geração de suportes de itens para a avaliação internacional PIRLS, voltadas para alunos do 4º ano do ensino fundamental, no qual utilizaram o modelo GPT-3 (Text-davinci-002) com as técnicas *zero-shot* e *few-shot*. O estudo demonstrou que, com o uso de *prompts* cuidadosamente elaborados, o modelo foi capaz de produzir textos com variedade temática, coesão e clareza comparáveis a materiais desenvolvidos por especialistas.

Embora trabalhos como [Lin and Chen 2024, Scaria et al. 2024, Wang et al. 2024, Lee et al. 2024] tenham feito a avaliação dos itens gerados por especialistas, e trabalhos como [Rockembach and Thom 2024, Zimerman et al. 2024] tenham utilizado a avaliação humana no contexto da AIG em português brasileiro; ainda não existem critérios consolidados que guiem a avaliação da qualidade dos itens em contextos de larga escala no Brasil. Além disso, não foram encontrados trabalhos que avaliam os resultados por meio de especialistas em língua portuguesa ou da área de

educação, segundo as diretrizes nacionais. Também não foram encontrados trabalhos que estudam a AIG no contexto de avaliação em larga escala em língua portuguesa.

Este trabalho abrange abordagens de uso de *templates* de itens com técnicas de engenharia de prompt, utilizando as versões mais recentes e estáveis do GPT-4.1, Gemini 2.0 flash e Deepseek v3024. Até o presente momento deste trabalho, não foram encontrados estudos que analisaram essas abordagens para geração de itens em língua portuguesa, utilizando critérios definidos para itens de alta qualidade.

4. Materiais e método

A Figura 2 apresenta as etapas adotadas por esse estudo. A geração do item é iniciada com a seleção da abordagem a ser utilizada e a definição das técnicas; esse passo também inclui a seleção do modelo. É realizada uma triagem a partir dos itens gerados. Todos esses passos vão ser melhor apresentados durante esta seção.

Figura 2. Fluxograma da metodologia

4.1. Seleção de abordagem

Foram definidos a experimentação com a geração de itens utilizando habilidades alinhadas à Base Nacional Comum Curricular (BNCC) através de três abordagens: (i) geração de itens sem fornecimento de suporte, (ii) geração dos itens com fornecimento de suporte e (iii) geração de itens de maior complexidade.

4.1.1. Geração de itens sem fornecimento de texto suporte

O objetivo da primeira abordagem é avaliar a capacidade das LLMs na geração do texto suporte (SLLM), do gênero carta, e dos demais elementos do item, como comando, gabarito, distratores e justificativas para cada alternativa. Essa abordagem trabalha com a geração de itens associados à habilidade: reconhecer relações lógico-discursivas de causalidade. Devem ser verificados critérios como qualidade, nível de dificuldade, público-alvo e compatibilidade do assunto. Os itens são gerados a partir de uma instrução que define as principais características do texto. A seguir, é apresentada uma instrução exemplo gerada por especialistas de língua portuguesa.

> Apresentar uma curiosidade científica, de até 12 linhas, com uma sintaxe que contenha períodos curtos, com poucas intercalações, formados por, no máximo, duas linhas, com léxico próprio do discurso formal com a presença de termos especializados, porém com estratégias de facilitação do discurso como a repetição de um elemento que seja central à compreensão do texto e outras pistas que permitam ao estudante lidar com um termo desconhecido. Quanto ao tema, deve-se privilegiar textos que tragam elementos do cotidiano das crianças, como informações sobre animais próximos ou até mesmo distantes de seu convívio, plantas, planetas, funcionamento do corpo humano, entre outros. No que se refere ao gênero, deve-se procurar textos de estrutura canônica, ou seja, aqueles em que, preferencialmente, o tema seja introduzido com uma pergunta no título e os parágrafos respondam à pergunta com exemplos que expliquem o tema aproximando-o do universo infantil.

A instrução delimita as possibilidades para que o texto de suporte gerado contenha as características que são consideradas importantes à um especialista. Todos os outros elementos de um item também possuem instruções que indicam suas características.

4.1.2. Geração do item com fornecimento de texto suporte

A segunda abordagem tem como objetivo avaliar a capacidade da LLM, dado um suporte gerado por humano (SH), de criar todos os elementos restantes do item, ou seja, comando, gabarito, distratores e justificativas. Assim como no caso anterior, essa abordagem trabalha com a geração de itens associados à habilidade: reconhecer relações lógico-discursivas de causalidade. A motivação é gerar itens semelhantes e possivelmente comparáveis. Novamente, cada uma dessas partes é gerada por uma instrução que define e delimita as características que devem ser apresentadas para cada um dos elementos do item. A seguir, um exemplo de instrução gerada por especialistas em língua portuguesa, o qual apresenta as características que o distrator deve possuir, delimitando os limites desse elemento:

> Os distratores devem apresentar informações que estejam em outras partes do texto, preferencialmente que não estejam no primeiro parágrafo ou na moral da história. Os distratores podem ser organizados: em ordem alfabética, em ordem alfabética inversa ou, ainda, com as informações que se apresentam na ordem em que aparecem no texto.

4.1.3. Geração de itens de maior complexidade

Complementando o experimento, a terceira abordagem tem como objetivo avaliar a capacidade das LLMs de gerar itens associados a habilidades mais complexas de leitura e interpretação dado um suporte gerado por humanos (SHHC). Especificamente, essa abordagem lida com itens que envolvam a habilidade de identificar relações lógico-discursivas de condição, marcadas por conjunções presentes em notícias. Nesta etapa, o texto suporte utilizado é fornecido ao modelo e são classificados como de nível 2, indicando uma maior complexidade intrínseca. O texto a seguir é um exemplo de instruçãopara a criação de um comando que solicite ao estudante identificar a relação de casualidade.

> Solicitar, por meio de um comando claro e objetivo, que o estudante identifique uma relação de causalidade marcada por conjunções mais usuais (porque, pois, já que, por isso). Esse comando deve trazer de forma explícita a consequência do fato (causa) narrado no texto. Direcionar a pergunta considerando o trecho do texto que a conjunção aparece, deixando-a expressa no comando.

Já no próximo exemplo, o item requer um raciocínio mais elaborado, pois o comando tende a ser mais implícito, o texto suporte é mais denso, e o reconhecimento da relação condicional exige uma interpretação. Dessa forma, esta abordagem representa um passo crucial ao direcionar o foco para a geração de itens de maior complexidade. A ênfase na identificação de relações condicionais, com base em textos mais elaborados e comandos que exigem maior inferência, visa testar a capacidade interpretativa das LLMs no desenvolvimento de questões avaliativas alinhadas às diretrizes da BNCC.

> Solicitar, por meio de um comando claro e objetivo, o reconhecimento do tipo de relação semântica estabelecida por elementos de coesão(condição) no texto. Pode ser por meio de uma pergunta contextualizada a ser respondida ou uma frase a ser completada pelo gabarito.

4.2. Combinação de Técnicas

Além das abordagens citadas na Subseção 4.1, uma variação de modelos de LLM e técnicas voltadas à engenharia de *prompt* também são aplicadas com o objetivo de melhorar o resultado gerado por parte das LLMs. A escolha dos modelos tem importância fundamental para a qualidade da solução de AIG produzida. Neste trabalho, foram escolhidos os modelos GPT-4.1, Gemini 2.0 Flash e DeepSeek V3 0324, por conta de serem as maiores versões estáveis de propósito geral de 3 grandes famílias de LLMs disponíveis durante o desenvolvimento do trabalho. O GPT é o modelo mais usado em trabalhos de AIG [Song et al. 2025], DeepSeek é o maior modelo de código aberto que tem mostrado resultados promissores para soluções de problemas de PLN e o Gemini tem demonstrado resultados interessantes em testes padronizados de *benchmark* [Team et al. 2023].

As técnicas mais comuns utilizadas em soluções de engenharia de *prompt* são o ***zero-shot*** e o ***few-shots***. O *few-shot* acontece quando o *prompt* é alimentado com um ou mais exemplos. O modelo ”entende”o padrão e desenvolve uma solução baseada nesses exemplos. O *zero-shot* é quando não há um exemplo e o *one-shot* é uma variação específica do *few-shots* que utiliza apenas um exemplo [Brown et al. 2020]. Também são utilizadas as técnicas ***Chain of Thoughts (CoT) - zero-shot***, que envolvem solicitar dentro do *prompt* para que a LLM detalhe passo a passo a geração da resposta [Kojima et al. 2023]; e a técnica ***Emotional prompt (EP)***, que envolve acrescentar uma ou mais frases de apelo emocional no *prompt*. Essas técnicas foram escolhidas por terem o potencial de melhorar a capacidade do modelo de gerar melhores resultados. Além disso, para cada caso, foi executado cinco vezes, totalizando 180 itens.

4.3. Triagem

Após a geração dos itens para cada técnica de *prompt* e tipo de item, foi realizada uma avaliação humana para verificar sua validade. Com o objetivo de reduzir o volume de itens encaminhados aos especialistas em avaliação em larga escala, foi definido um conjunto de critérios claros e de fácil aplicação. Dessa forma, nesta etapa, os itens passaram por uma triagem, sendo enviados aos especialistas apenas aqueles que apresentaram uma qualidade mínima aceitável. A triagem foi realizada por dois graduandos que possuem conhecimento básico sobre geração de itens. Os critérios (Tabela 1) utilizados foram elaborados com base nas diretrizes apresentadas na Seção 2 de forma a ter respostas binárias.

Código	Descrição
E1	Item apresenta erro gramatical
E2	Item apresenta erro semântico
E3	Item apresenta alinhamento com a competência avaliada
E4	Item apresenta texto de suporte e comando compatíveis com a instrução
E5	Item apresenta gabarito incorreto ou inválidos
E6	Item apresenta distratores não plausíveis ou inválidos
E7	Item apresenta justificativas de distratores e gabarito não plausíveis/ inválidos

Tabela 1. Critérios de exclusão

A avaliação da gramática (E1) e da semântica do item (E2) verifica se os itens estão corretamente escritos e se a mensagem transmitida faz sentido. O critério E3 avalia

o alinhamento do item gerado com a competência definida nas instruções do *prompt*. Em outras palavras, verifica-se se os itens foram elaborados de acordo com os aspectos de dificuldade, forma e demanda cognitiva esperada. O critério E4 examina a validade do texto de suporte e do comando gerados, considerando se estão de acordo com as instruções, se são adequados ao contexto e se não representam cópias do suporte apresentado nos exemplos, especialmente no caso da técnica *few-shot*. Já o critério E5 avalia o gabarito, verificando se ele está correto, responde adequadamente ao comando e é único. O critério E6 avalia a validade dos distratores, verificando se eles respondem ao comando, se poderiam ser escolhidos por um aluno que não domine a competência avaliada e se apresentam paralelismo — ou seja, se possuem estruturas semelhantes em termos de tamanho e sintaxe, evitando que alguma alternativa se destaque indevidamente e facilite a identificação do gabarito ou induza ao erro. Por fim, o critério E7 examina as justificativas dos distratores e do gabarito, observando se há lógica e coerência nas explicações fornecidas.

Durante o processo de triagem, foram eliminados todos os itens considerados duplicados ou incompletos, a fim de garantir a consistência e a qualidade do conjunto final. Após essa etapa, restaram 59 itens válidos. Para compor o conjunto utilizado na etapa de avaliação especializada, foi realizada uma seleção aleatória, restringindo-se a um item por combinação de variáveis (modelo, abordagem e técnica). Esse procedimento resultou em um total de 25 itens únicos, representativos das diferentes configurações experimentais.

4.4. Avaliação por especialistas em itens para larga escala

Após a triagem inicial dos itens, dois especialistas em geração de itens de língua portuguesa avaliaram os resultados, visando garantir a qualidade linguística e semântica dos conteúdos gerados. Os especialistas possuem vínculo de pesquisadores em produção de itens no Centro de Políticas Públicas e Avaliação da Educação (CAEd). Os critérios originalmente definidos na subseção 4.3 foram reapresentados e adaptados para compor o conjunto de critérios de avaliação adotados nesta etapa. Além disso, foram incorporados novos objetivos específicos para abranger aspectos essenciais na composição de itens destinados a avaliações em larga escala, ampliando o rigor e a abrangência da avaliação.

Código	Descrição
Q1	Suporte está adequado?
Q2	Comando reflete a habilidade?
Q3	Gabarito está correto?
Q4	Gabarito é único?
Q5	Existe paralelismo entre as alternativas?
Q6	Os distratores estão plausíveis?
Q7	Justificativas estão coerentes?
Q8	Alternativas apresentam elementos que não estão no suporte?
Q9	Item está adequado em termos de alinhamento com a habilidade?
Q10	Item poderia ser usado em uma avaliação com pouca ou nenhuma alteração?

Tabela 2. Rúbrica de avaliação

Os critérios de avaliação (Tabela 2) foram organizados para verificar diferentes dimensões da qualidade dos itens. O primeiro critério (Q1) diz respeito à adequabilidade do suporte fornecido, assegurando que o material de apoio apresentado esteja coerente

e consistente com as instruções e o contexto do item. O segundo critério (Q2) avalia a validade do comando, isto é, se a solicitação feita ao respondente está clara, precisa e sem ambiguidades. Já o terceiro e quarto critério (Q3 e Q4) referem-se, respectivamente, à correção e unicidade do gabarito, garantindo que a resposta correta esteja correta de fato e que exista uma única alternativa correta para evitar confusão ou múltiplas interpretações. Outro ponto importante analisado foi o paralelismo das alternativas (Q5), que nesse caso é analisado como um critério à parte em relação aos distratores e ao gabarito.

No que tange aos distratores, o critério Q6 avalia a plausibilidade dessas alternativas incorretas. Para isso, verifica-se se os distratores respondem adequadamente ao comando, se poderiam ser escolhidos por um respondente que não domina a habilidade avaliada e se apresentam coerência lógica, assegurando que não sejam obviamente descartáveis. Por sua vez, o critério Q7 refere-se à coerência das justificativas apresentadas para os distratores e para o gabarito, observando se há uma fundamentação lógica e consistente que esclareça por que uma alternativa está correta ou incorreta.

Além desses aspectos, a avaliação contemplou outras dimensões importantes para a qualidade dos itens. A qualidade das alternativas (Q8) foi verificada, especialmente quanto à presença de elementos não contemplados no suporte, que poderiam gerar dúvidas ou enviesar a resposta. Também se avaliou o alinhamento do item com a habilidade ou competência a ser avaliada (Q9), certificando que o conteúdo corresponde aos objetivos pedagógicos. Por fim, analisou-se a qualidade do item quanto à aplicabilidade prática em avaliações, considerando sua utilização com pouca ou nenhuma modificação pelas especialistas (Q10), ou seja, o item está adequado ou próximo de satisfazer os critérios para avaliações em larga escala.

5. Resultados

Nesta seção são discutidos os resultados da triagem e dos especialistas, além das diferenças, concordâncias e discordâncias entre essas duas avaliações. O código-fonte, junto com os *prompts* e os itens gerados neste trabalho, foram disponibilizados em um repositório público[1]. É importante apontar que as análises foram realizadas em uma quantidade não tão expressiva devido a restrições de tempo. Portanto, os resultados apresentados são relevantes, mas fornecem indícios e não constituem uma conclusão absoluta.

5.1. Resultados da triagem

A Figura 3 apresenta a distribuição relativa dos critérios de exclusão apresentados na Seção 4.3, considerando tanto ocorrências de critérios únicos quanto as combinações de critérios na exclusão de itens. Cada segmento do gráfico representa a frequência de determinado critério, ou combinação de critérios, que foram identificados nos itens gerados.

Observa-se que o critério E6 individualmente foi o mais recorrente, representando 60,3% das exclusões. Além disso, esse critério também esteve presente em diversas combinações com outros critérios, o que evidencia sua ampla incidência no processo de triagem. O critério foi aplicado, principalmente, em situações em que os distratores apresentavam conteúdo absurdo ou incompatível com o comando da questão, comprometendo a função avaliativa do item ao facilitar a identificação da resposta correta por estudantes

[1] https://github.com/ufjf-items/geracao-itens-sbie-2025.git

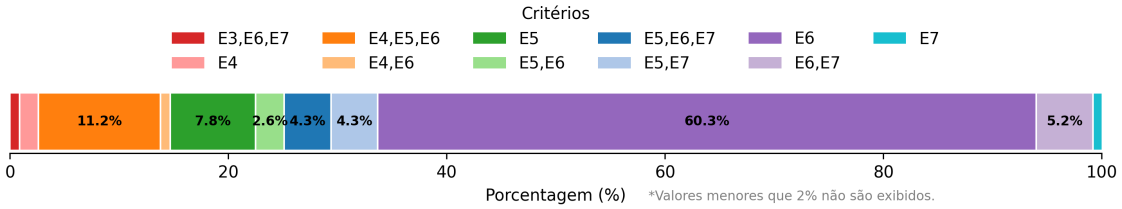

Figura 3. Proporção de critério de exclusão aplicados na triagem

que não dominam a habilidade avaliada. Em contraste, foi percebido que o critério de alinhamento com a competência avaliada (E3) teve ocorrência baixa, estando presente apenas em combinações pouco frequentes. Esse padrão indica que os modelos utilizados tendem a gerar itens com temas coerentes com a habilidade desejada, mas falham na elaboração de distratores plausíveis, característica fundamental para itens de qualidade.

A Figura 4 mostra a proporção dos itens aceitos agrupando-os pela LLM utilizada, pelo suporte utilizado ou pelo tipo de *prompt* (técnica de geração). O modelo DeepSeek teve desempenho levemente inferior aos demais. Em termos de abordagem, a habilidade mais complexa obteve os piores resultados, enquanto o suporte humano simples apresentou melhor desempenho. Quanto às técnicas, a CoT+EP ficou ligeiramente abaixo das demais. Esse resultado contraria a expectativa de que estratégias mais elaboradas, como CoT+EP, contribuiriam para uma maior taxa de aceitação, sugerindo que a complexidade adicionada ao *prompt* pode não ter sido eficaz nesse contexto.

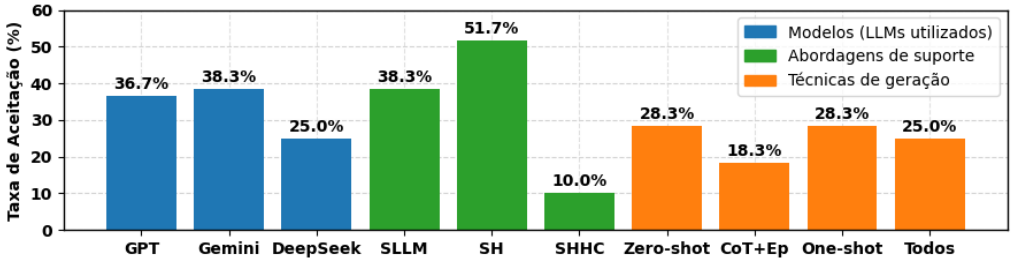

Figura 4. Proporção de itens aceitos pela triagem

5.2. Resultados dos especialistas

Os especialistas analisaram os itens com o objetivo de responder à seguinte pergunta: "O item poderia ser usado em uma avaliação com pouca ou nenhuma alteração?" (critério Q10). Foram avaliados os 25 itens que passaram pela triagem, sem que os avaliadores tivessem conhecimento prévio sobre sua origem ou técnica de geração. A Tabela 3 apresenta a distribuição das respostas (Sim, Não) e o total avaliado segundo três perspectivas definidas posteriormente para análise das avaliações: modelo de linguagem utilizado, abordagem adotada e técnica de geração. O destaque ao critério Q10 se justifica por ele refletir diretamente a qualidade do item gerado.

Na análise dos resultados, observa-se que o modelo GPT apresentou o menor número de itens aprovados (3 de 9), indicando desempenho inferior sob a ótica dos especialistas, especialmente comparado ao modelo Gemini (8 de 9) e DeepSeek (6 de 7). Essa

	Modelos			Abordagens			Técnicas			
Critérios	GPT	Gemini	DeepSeek	SLLM	SH	SHHC	Baseline	CoT	One-shot	Todos
Sim	3	8	6	7	9	1	5	4	4	4
Não	6	1	1	2	3	3	2	2	1	3
Total	9	9	7	9	12	4	7	6	5	7

Tabela 3. Contagem de itens aceitos, rejeitados e total por categoria

diferença sugere que, apesar dos textos gerados pelo GPT poderem parecer adequados em triagem inicial, apresentam falhas mais evidentes quando avaliados com rigor técnico.

No que se refere às abordagens, os itens gerados por SH (9 aprovações em 12) e por SLLM (7 de 9), ambos voltados ao 3º ano do Ensino Fundamental, obtiveram as melhores taxas de aprovação, sendo considerados os mais adequados pelos especialistas. Em contraste, a abordagem SHHC apresentou desempenho significativamente inferior, com apenas 1 aprovação em 4 itens avaliados. Essa baixa adequação repete os resultados observados na etapa de triagem, que já havia indicado dificuldades das LLMs em lidar com conteúdos de maior complexidade. Apesar de poucos itens dessa categoria terem sido encaminhados para avaliação especializada, o fato de apenas um ter sido aceito reforça a limitação da abordagem para esse tipo de item. Em relação às técnicas, os resultados se mostraram mais equilibrados, com leve destaque para a técnica one-shot (4 de 5 itens aprovados), embora a diferença em relação às demais técnicas (como baseline e CoT+EP) seja pequena e não estatisticamente significativa. Portanto, ainda que haja variações entre os métodos utilizados, os dados reforçam a importância da avaliação qualitativa especializada para validar o uso pedagógico dos itens gerados.

A Figura 5 relaciona, dentre os 25 itens selecionados, qual foi o resultado em relação a cada critério analisado pelos especialistas conforme apresentado na Tabela 2. Nessa representação, cada barra representa a proporção de itens que receberam resposta "Sim" ou "Não" para um determinado critério. A resposta "Sim" indica que o item atendeu ao critério avaliado, já a resposta "Não" aponta que o item apresentou falhas em relação ao critério avaliado. Dentro desse resultado, é possível perceber pela Figura 5 que os critérios de adequação com o suporte (Q1), corretude (Q3) e unicidade do gabarito (Q4) e se as alternativas apresentam elementos que não estão no suporte (Q8) tiveram maiores concordâncias com a triagem, visto que poucos itens que passaram pela triagem foram reprovados pelos especialistas com base nesses critérios. Isso indica que a triagem teve mais êxito em fazer a filtragem dos itens com base nesses critérios.

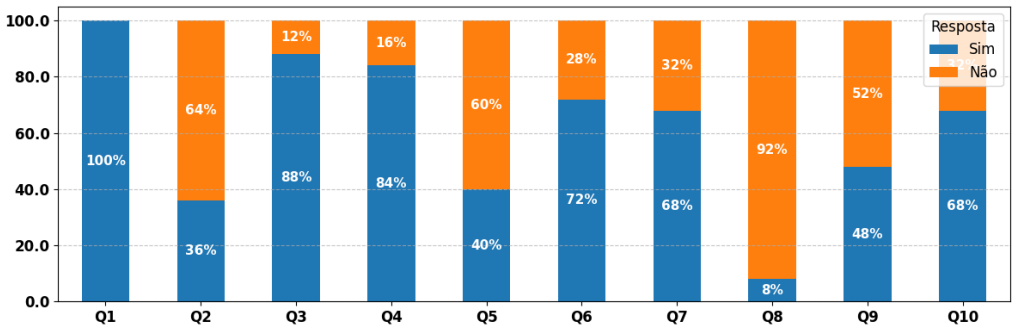

Figura 5. Porcentagem dos resultados da avaliação dos especialistas

Por outro lado, percebe-se que as questões sobre o comando (Q2) e o alinhamento (Q9) tiveram maiores discordâncias em relação à triagem. Enquanto poucos itens foram excluídos na triagem por conta do comando (E4) e menos de 1% dos itens foram excluídos por conta do alinhamento (E3), mais da metade dos itens avaliados pelos especialistas não estavam adequados em termos de alinhamento (Q9) e do comando (Q4). Isso reflete uma desconformidade em relação à avaliação da triagem, indicando que avaliações realizadas por não especialistas possuem maiores limitações quanto à avaliação dos critérios.

6. Conclusão

Este estudo apresenta um experimento com avaliação humana aplicada a um recorte específico e, portanto, os resultados devem ser interpretados como indícios observados em um contexto delimitado, não como conclusões universais. Um volume maior de avaliações permitiria alcançar significância estatística, mas isso exigiria um esforço considerável de uma equipe especializada. Também é importante considerar a complexidade dos LLMs, a variação gerada por diferentes *prompts* e a imprevisibilidade das respostas. Para reduzir essa variabilidade, foram utilizadas configurações determinísticas, com o objetivo de aumentar a consistência dos resultados.

Apesar de seu caráter exploratório, o estudo observou tendências quanto a pontos fortes e limitações das técnicas de AIG baseadas em modelos generativos. Essas observações sugerem oportunidades de pesquisa para o aprimoramento dos mecanismos de geração, buscando maior eficiência e aderência aos critérios de qualidade exigidos em avaliações em larga escala. Na triagem e na análise especializada, foi verificada a falta de paralelismo das respostas; a falta de conexidade entre gabaritos, distratores e comandos; a aplicação não usual da língua e, por fim, itens que não avaliam a habilidade desejada. Esses itens, passando pela triagem, mas não pelas avaliadoras, reiteram a complexidade do trabalho de avaliação e evidenciam a importância da avaliação humana ser feita por especialistas em contraste com os trabalhos de AIG voltados à língua portuguesa feitos até o momento. LLMs são ferramentas com alta capacidade de seguir padrões e demonstraram desempenho promissor na geração de suportes do tipo cartas para os modelos selecionados. Quanto à dificuldade das questões, em alguns casos foram produzidas questões complexas para a faixa etária definida, revelando a limitação das LLMs em garantir alinhamento adequado das questões e indicando a necessidade de pesquisas futuras sobre a adequação da dificuldade dos itens.

Diante dessas limitações, a análise dos resultados e o andamento do estudo indicam lacunas que podem orientar pesquisas futuras. A utilização de técnicas mais avançadas de *prompt engineering*, como RAG (*Retrieval Augmented Generation*) e arquiteturas multiagentes, podem ser particularmente promissora. Além disso, o uso de *fine-tuning* e *prompt-tuning* em modelos de LLM pode contribuir para especializá-los na geração de itens educacionais em português brasileiro. Outra linha de investigação promissora consiste na avaliação de itens gerados mediante a outras estratégias de avaliação, como a mensuração da dificuldade, conforme proposto em [Amorim et al. 2019], ou a utilização do teste de Turing [Song et al. 2025], que podem complementar a análise feita por especialistas e enriquecer a compreensão sobre a performance dos modelos.

Referências

Amorim, M., Simões, J., Assis, F., Pinheiro, J., Menasch, D., Motta, C., and Pacheco, A. (2019). Aumentando a interatividade no ensino a distância via geração automática de questões: Desafios, soluções via aprendizado por máquina e um estudo de caso no cederj. In *Anais do XXVII Workshop sobre Educação em Computação*, pages 188–202, Porto Alegre, RS, Brasil. SBC.

Awalurahman, H. W. and Budi, I. (2024). Automatic distractor generation in multiple-choice questions: a systematic literature review. *PeerJ Computer Science*, 10:e2441.

Bezirhan, U. and von Davier, M. (2023). Automated reading passage generation with openai's large language model. *Computers and Education: Artificial Intelligence*, 5:100161.

Bollela, V. R., Borges, M. d. C., and Troncon, L. E. d. A. (2018). title = Avaliação Somativa de Habilidades Cognitivas: Experiência Envolvendo Boas Práticas para a Elaboração de Testes de Múltipla Escolha e a Composição de Exames,. *Revista Brasileira de Educação Médica*, 42:74 – 85.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Circi, R., Hicks, J., and Sikali, E. (2023). Automatic item generation: foundations and machine learning-based approaches for assessments. In *Frontiers in Education*, volume 8, page 858273. Frontiers Media SA.

Gierl, M. J. and Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Applied Psychological Measurement*, 42(1):42–57. PMID: 29881111.

Gierl, M. J., Lai, H., and Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge.

Greaney, V. and Kellaghan, T. (2008). *Assessing national achievement levels in education*, volume 1. World Bank Publications.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Routledge.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2023). Large language models are zero-shot reasoners.

Lee, J., Smith, D., Woodhead, S., and Lan, A. (2024). Math multiple choice question generation via human-large language model collaboration. *arXiv preprint arXiv:2405.00864*.

Lin, Z. and Chen, H. (2024). Investigating the capability of chatgpt for generating multiple-choice reading comprehension items. *System*, 123:103344.

Maity, S., Deroy, A., and Sarkar, S. (2024). Investigating large language models for prompt-based open-ended question generation in the technical domain. *SN Computer Science*, 5(8):1128.

Palacios, C. and de Oliveira, L. K. M. (2022). *Avaliação da educação básica e seus instrumentos*. Carlos Palacios Carvalho da Cunha e Melo.

Richichi, R. V. (1996). An analysis of test bank multiple-choice items using item response theory.

Rockembach, G. and Thom, L. (2024). Investigating the use of intelligent tutors based on large language models: Automated generation of business process management questions using the revised bloom's taxonomy. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 1587–1601, Porto Alegre, RS, Brasil. SBC.

Scaria, N., Dharani Chenna, S., and Subramani, D. (2024). Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In Olney, A. M., Chounta, I.-A., Liu, Z., Santos, O. C., and Bittencourt, I. I., editors, *Artificial Intelligence in Education*, pages 165–179, Cham. Springer Nature Switzerland.

Setiawan, H., Hidayah, I., and Kusumawardani, S. S. (2022). Automatic item generation with reading passages: A systematic literature review. In *2022 8th International Conference on Education and Technology (ICET)*, pages 250–255.

Silva, M. M. d. S., Reihn, C., Soares, A., and Soares, T. M. (2016). A abordagem da avaliação educacional em larga escala nos cursos de graduação em pedagogia. *Revista Brasileira de Estudos Pedagógicos*, 97(245):46–67.

Song, Y., Du, J., and Zheng, Q. (2025). Automatic item generation for educational assessments: a systematic literature review. *Interactive Learning Environments*, pages 1–20.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wang, L., Song, R., Guo, W., and Yang, H. (2024). Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*, pages 1–26.

Zimerman, F., Duarte, F. H., Silva, P. H., and Fortes, R. (2024). Explorando chatgpt para criação automática de questões práticas de programação de computadores. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 2353–2364, Porto Alegre, RS, Brasil. SBC.