

ELABORAÇÃO DE ITENS DE MATEMÁTICA COM AUXÍLIO DE INTELIGÊNCIA ARTIFICIAL GENERATIVA

Preparation of mathematics items with the aid of generative artificial intelligence

Andrey Camurça da Silva¹
Mario Tanaka Filho²

Resumo: Este artigo explora a viabilidade de utilizar a Inteligência Artificial (IA) generativa - modelos que têm a capacidade de gerar conteúdo inédito a partir de dados pré-existentes - no processo de criação e revisão de itens para avaliação do aprendizado em matemática, a qual pode se tornar mais um recurso de auxílio ao professor. Para isso, propõe-se um modelo básico de comando textual que fornece orientações gerais para a elaboração de itens, a partir da descrição da habilidade a ser avaliada pelo item e de um contexto que o item deve abordar. Alguns dos itens gerados, juntamente com itens elaborados por especialistas, foram respondidos por um grupo de estudantes e posteriormente submetidos a uma análise de parâmetros como poder discriminativo, consistência interna e nível de dificuldade. O estudo revelou limitações em alguns itens, especificamente no comando para resposta e nas alternativas, apresentando também ausência de gabarito em alguns casos. No entanto, alguns itens gerados exibiram um poder discriminativo semelhante ao dos itens produzidos por especialistas, destacando o potencial desses recursos como auxiliares no processo de elaboração de itens.

Palavras-chave: inteligência artificial; avaliação; educação matemática; elaboração de itens; chats inteligentes.

Abstract. *This article explores the feasibility of using generative Artificial Intelligence (AI) - models that have the ability to generate novel content from pre-existing data - in the process of creating and reviewing items for the assessment of learning in mathematics, which could become an additional resource to aid the teacher. For this, a basic text command model is proposed that provides general guidelines for the elaboration of items, based on the description of the skill to be assessed by the item and a context that the item should address. Some of the generated items, along with items developed by experts, were answered by a group of students and subsequently subjected to an analysis of parameters such as discriminative power, internal consistency, and difficulty level. The study revealed limitations in some items, specifically in the command for response and in the alternatives, also presenting the absence of an answer key in some cases. However,*

¹ Mestre em Matemática pela Universidade Federal do Oeste do Pará - UFOPA. E-mail: andreycamurca@gmail.com

² Professor Associado I da Universidade Federal do Oeste do Pará - UFOPA. E-mail: tanakafi@gmail.com

some generated items exhibited a discriminative power similar to the items produced by experts, highlighting the potential of these resources as aids in the item elaboration process.

Keywords: *artificial intelligence; assessment; mathematics education; elaboration of questions; smart chats.*

1. Introdução

A Inteligência Artificial (IA) é um ramo da ciência da computação que busca criar sistemas capazes de simular comportamentos e atributos essencialmente humanos, como o aprendizado, o raciocínio, a percepção e a comunicação (Delipetrev; Tsinarakii; Kostić, 2020; Russell; Norvig, 2013; Samoili et al., 2021). Embora não se saiba com exatidão quando surgiram os primeiros protótipos de IA, Haenlein e Kaplan (2019) atribuem sua origem à década de 1940. Esse período foi marcado pelo desenvolvimento de uma máquina que decifrou o código Enigma usado pelo exército alemão na 2ª Guerra Mundial e, em seguida, pela proposta de um teste para avaliar se uma máquina é ou não inteligente, conhecido como Teste de Turing. A partir daí, a IA passou por diversas fases de avanços e retrocessos, marcadas por diferentes abordagens e paradigmas (Buchanan, 2005; Haenlein; Kaplan, 2019).

Um dos avanços relativamente recentes na área de IA foi o desenvolvimento de Grandes Modelos de Linguagem (GML), que são redes neurais artificiais treinadas com expressiva quantidade de dados textuais, capazes de produzir, a partir de um estímulo inicial, textos coerentes e consistentes sobre diversos assuntos (Sejnowski, 2023; Yu, 2023). Um exemplo de GML é o GPT 3.5 desenvolvido pela OpenAI, que pode ser acessado com auxílio do ChatGPT – interface conversacional que permitem ao usuário interagir com o modelo de linguagem por meio de texto. Assim, o ChatGPT faz parte do grupo dos chats inteligentes utilizados para realizar, entre outras tarefas, tradução, síntese, revisão de textos e produção de novos textos.

Considerando o potencial dessas ferramentas, a IA pode trazer desafios e possibilidades para o campo da educação, especialmente no que se refere à avaliação da aprendizagem (Figueiredo et al., 2023; Silva, 2023). Na sala de aula, a avaliação é fundamental para o processo educativo, uma vez que permite mapear possíveis habilidades já desenvolvidas, bem como identificar as dificuldades e os avanços de cada educando (Luckesi, 2013, 2022; Sousa, 2000).

No âmbito das avaliações externas, parte da avaliação é conduzida com auxílio de testes cognitivos padronizados, cujo objetivo é constatar níveis de proficiências de determinados grupos (por escola, município ou UF) que são utilizados como um dos indicadores da qualidade educacional (INEP, 2018; Sousa, 2000). No entanto, para essa última finalidade, a elaboração de itens (questões) é uma tarefa que exige competência e critério por parte de quem os elabora, uma vez que envolve a seleção dos conteúdos e das habilidades preconizadas nas matrizes de referência, a formulação de situações-problema que estimulem o raciocínio e a criatividade dos educandos. Os itens, como

unidade mais elementar dos testes avaliativos, devem exibir características estatísticas que atestem sua capacidade discriminativa, nível de dificuldade, consistência interna, e até mesmo a atratividade de acerto ao acaso (Pasquali, 2017; Vianna, 1978).

Diante do exposto, este artigo discute a possibilidade de utilização de Inteligência Artificial (IA) generativa – modelos capazes de criar conteúdo novo a partir de dados existentes – no processo de elaboração e revisão de itens para avaliação da aprendizagem em matemática, a qual pode se tornar mais um recurso de auxílio ao professor. Este estudo é parte da dissertação de mestrado do autor do presente estudo (Silva, 2023), na qual foram empregados os recursos Bing Chat (atualmente conhecido como Copilot) e o ChatGPT na elaboração ou revisão de itens de múltipla escolha que compuseram um teste de matemática. O teste foi respondido por um grupo de 61 estudantes do 9º ano do ensino fundamental de uma escola pública no interior da Amazônia.

Ademais, o presente artigo apresenta uma seção que traz reflexões sobre o uso de inteligência artificial na educação, especialmente no âmbito da avaliação da aprendizagem. Em seguida, discute aspectos metodológicos, apresentando um modelo de conversação para a criação de itens de múltipla escolha a partir de comandos textuais. Na seção de resultados, alguns itens produzidos são analisados quanto ao nível de dificuldade, poder discriminativo e consistência interna, assim como em relação ao conteúdo do item. Por fim, na última seção, o artigo discute alguns desafios e possibilidades da IA generativa no processo de elaboração de itens para avaliação da aprendizagem em matemática no ensino fundamental.

2. Inteligência Artificial: um recurso para a avaliação da aprendizagem

No âmbito da educação, a IA é um campo que se desenvolve desde a década de 1970, com sistemas de tutoria inteligentes e autoria baseados em redes semânticas, sistemas especialistas e aprendizado de máquina. Algumas soluções de IA visavam promover a aprendizagem dos alunos por meio de modelos explícitos do conteúdo, da didática e do conhecimento do aluno, adaptando-se às suas necessidades e características individuais, o que a torna promissora para a personalização do ensino (Boulay, 2022; Giraffa; Khols-Santos, 2023).

Nesse contexto, o desenvolvimento de sistemas de IA impulsionados pelas pesquisas em Aprendizado de Máquina (AM) trouxeram soluções para diversos setores da educação. Com efeito, Gatti (2019) aponta que pesquisas sobre aplicações da IA na educação vem sendo desenvolvidas no Brasil, as quais englobam todos os níveis de ensino. Tais pesquisas tem como foco ensino a distância, gestão, personalização, avaliação e inclusão, sendo mais frequentes no ensino superior do que na educação básica.

Ao mesmo tempo que busca superar desafios complexos dos processos de gestão e do ensino, a IA na educação é desafiada no campo da ética, especialmente no que se refere à privacidade, ao uso dos dados dos alunos e às tomadas de decisões por IA

pautadas em dados que reproduzem, em algum nível, discriminações e injustiças (Teles; Nagumo, 2023; Yu, 2023). Algumas soluções de IA são empregadas na seleção de cursos, atribuição de notas, avaliação, recomendação de carreiras, produção de material e tutorias personalizadas. Essas decisões devem ser éticas, justas e equitativas, evitando a reprodução de preconceitos ou desigualdades presentes nos dados usados pela IA. Assim, é necessário proteger a privacidade e os direitos dos educandos por meio de leis e regulamentos adequados (Boulay, 2022).

Os debates sobre IA segura na educação resultou na realização da Conferência Internacional sobre Inteligência Artificial e Educação (UNESCO, 2019). O evento realizado em 2019 reuniu autoridades governamentais e diversos atores sociais em Pequim, na China, para debater o papel da IA na educação. O evento resultou em 10 recomendações de ações para os governos e as partes interessadas nessa área. A recomendação de número 4 é voltada para IA para aprendizagem e avaliação da aprendizagem. Ela propõe usar a IA para personalizar a aprendizagem e avaliar o progresso dos alunos de forma eficiente e precisa.

Assim, o emprego da IA na avaliação da aprendizagem é uma possibilidade proeminente, mas que traz consigo adversidades. Tendo em vista que, na concepção de Luckesi (2013), avaliar consiste em um processo de coleta, análise e síntese de dados que visa estabelecer o valor ou a qualidade do objeto avaliado, mediante a confrontação com um critério de qualidade definido previamente, há de se considerar que tal ato é eticamente carregado e que precisa de transparência, bem como supervisão. Entende-se, aqui, que o papel da IA nesse processo deva se reduzir à natureza operacional, possibilitando celeridade e eficiência na avaliação, mas que seus critérios sejam rigorosamente estabelecidos e revisados por profissionais da educação, para assegurar que tais avaliações estejam em consonância com os valores e os objetivos educacionais preconizados no currículo.

O entendimento a respeito da avaliação da aprendizagem assumido no presente estudo é baseado em Luckesi (2013, 2022) e Sousa (2000). De acordo com Sousa (2000), a avaliação é definida em função do contexto em que está inserida. Na sala de aula, a avaliação é um instrumento para acompanhar e melhorar o processo de ensino e aprendizagem. Esta, além de envolver instrumentos quantitativos, como os testes cognitivos, precisa ser qualitativa, contínua e inclusiva, levando em conta não apenas o desempenho, mas também as habilidades, competências e atitudes dos alunos. Avaliar, nessa perspectiva, vai além da verificação, sendo esta última uma prática excludente e estática, que reduz a aprendizagem a uma coisa (Luckesi, 2013, 2022).

A avaliação de sistema, mais abrangente, visa subsidiar as políticas públicas educacionais em diferentes níveis de gestão. Analisa o sistema educacional como um todo, considerando seus componentes, estrutura, organização e resultados, buscando identificar e corrigir as desigualdades e injustiças (Luckesi, 2013, 2022; Sousa, 2000). A avaliação da aprendizagem conduzidas via testes cognitivos padronizados se insere na tentativa de mensurar os resultados educacionais em nível de sistema, e por essa razão

tem seus códigos e conceitos cada vez mais populares no meio escolar, o que insere tais instrumentos avaliativos no fazer pedagógico docente.

Com efeito, essa inserção pode trazer riscos de uma rotina exaustiva de testes cognitivos objetivos, os quais interferem no trabalho pedagógico e no currículo escolar (Gimenes et al., 2013). A isso se soma a pressão externa sobre os professores para que seus planejamentos atendam às matrizes de referência e aos modelos de itens dos testes, o que pode levar ao empobrecimento do processo avaliativo.

Assim, o uso de IA para avaliação deve tomar como perspectiva a diversificação dos instrumentos avaliativos e seu uso para o aperfeiçoamento do ensino, permitindo a criação de testes adequados e válidos, que possam auxiliar o docente nas tomadas de decisão para o aperfeiçoamento de suas práticas. Tales e Nagumo (2023) apontam a possibilidade de desenvolvimento de IA baseada em diferentes modelos pedagógicos, como os de Vygotsky e Paulo Freire. Esses modelos, que estimulam a interação e o trabalho em grupo, promovem uma educação produtiva e solidária, não se limitando ao desempenho individual dos alunos em testes, predominantes no modelo behaviorista.

Citam-se, aqui, alguns recursos baseados em IA úteis para a avaliação da aprendizagem. Dois deles são Bing Chat (atual Microsoft Copilot) e ChatGPT. Essas ferramentas, baseadas em Grandes Modelos de Linguagem (GML) e treinadas com bilhões de parâmetros, são capazes de produzir textos, resolver problemas e executar uma expressiva variedade de tarefas (Yu, 2023), o que possibilita dar celeridade à produção de conteúdo, planejamentos pedagógicos, instrumentos avaliativos e de materiais didáticos. Tendo em vista que o Bing Chat é baseado no modelo GPT 4.0, o enfoque das discussões seguintes será dado ao ChatGPT.

Com o objetivo de verificar como o ChatGPT pode ser usado na educação, tais como seu uso para melhorar o ensino e a aprendizagem, Lo (2023) conduziu uma revisão de 50 artigos sobre o tema nos primeiros meses de lançamento do GPT-3.5 no ano de 2022. O estudo aponta que o recurso foi um assistente promissor para professores e alunos. Para professores, o ChatGPT foi empregado para gerar materiais de curso, sugestões e traduções. Na avaliação, o ChatGPT foi empregado para gerar tarefas, exercícios, questionários, cenários e avaliar o desempenho dos alunos. Os alunos, por outro lado, usaram o recurso para tutoria virtual e revisão de trabalhos escritos.

No entanto, a qualidade do conteúdo gerado e a desconfiança quanto à autoria levantam preocupações, revelando a necessidade de qualificação de docentes e de orientações aos alunos direcionadas para o uso confiável e ético da IA. Assim, qualidade, confiabilidade, ética e segurança dos dados são, portanto, questões que devem ser consideradas na aplicação de IA em todos os campos, sobretudo no campo da educação (Giraffa; Khols-Santos, 2023; Lo, 2023; Yu, 2023).

No âmbito da educação matemática, o ChatGPT é uma ferramenta de apoio ao ensino e à aprendizagem que proporciona, por meio de diálogos naturais e personalizados, interações entre alunos, professores e conteúdos matemáticos, permitindo confrontações e análises das limitações e potencialidade da ferramenta. Entre

os benefícios, estão a motivação, o *feedback*, a autonomia e a diversidade de abordagens (Santos; Sant'Ana; Sant'Ana, 2023).

Um outro recurso digital relevante para avaliação da aprendizagem – o *Quizizz* – é voltado para a gamificação e mais recentemente vem incorporando IA para gerar conteúdo e avaliar a aprendizagem. O recurso permite que o professor crie itens (questões) a partir de comandos textuais, links de sites ou de vídeos da internet. Isso possibilita a criação de avaliações e atividades gamificadas, além de propiciar relatórios detalhados sobre o nível de aprendizado de cada indivíduo. Essa personalização do ensino é um instrumento alternativo para lidar com a heterogeneidade presente nas salas de aula, muitas delas lotadas, ao passo que otimiza o tempo e os recursos dos professores que pode contar com a automatização da correção de tarefas e da análise de dados de desempenho. Ferramentas como essa liberam tempo para que os docentes se concentrem em atividades de ensino e orientação individualizadas, o que pode impactar positivamente na qualidade do ensino.

Apesar das promessas e benefícios da IA na avaliação da aprendizagem, é fundamental ponderar sobre os desafios e obstáculos que podem comprometer sua efetividade e alcance. A disparidade no acesso à internet e às tecnologias digitais, especialmente em comunidades rurais e periféricas, configura-se como um impedimento para a inclusão digital e a equidade na educação. Políticas públicas direcionadas à democratização do acesso à tecnologia e à formação continuada de professores para o uso eficaz das ferramentas de IA são medidas essenciais para superar tais desafios e garantir que todos os alunos e docentes possam se beneficiar das inovações proporcionadas pela IA.

3. Metodologia

O estudo utiliza duas ferramentas de IA generativa, o ChatGPT, que usa o modelo GPT-3.5, e o Bing Chat, que tem uma relação de cooperação com o GPT-4 da OpenAI. Ambas as ferramentas usam Grandes Modelos de Linguagens para gerar textos, mas diferem em funcionalidades. O Bing Chat, um serviço gratuito da Microsoft, destaca-se pela capacidade de realizar buscas na web e apresentar resultados de forma visual e lógica. O ChatGPT, disponibilizado gratuitamente pela OpenAI, usa um modelo anterior ao GPT-4 e tem como funcionalidade gerar textos em linguagem natural sobre diversos domínios, embora tenha limitações, como a possibilidade de gerar informações incorretas e ter conhecimento limitado do mundo e eventos após 2021.

Um modelo básico de conversação para a elaboração de itens utilizando essas ferramentas começa com a definição de instruções em forma de texto sobre como o item deve ser elaborado, incluindo suas características básicas: formato (objetivo ou discursivo), número de alternativas, número de gabaritos e extensão das alternativas. As instruções têm como base as recomendações oriundas dos guias de elaboração de itens do Centro de Políticas Públicas e Avaliação da Educação (CAED) e do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) (INEP, 2010; CAED, 2008).

Nessa etapa, também se define um processo de produção de item que a ferramenta deve seguir. Dito isso, o usuário (professor/elaborador) insere, individualmente, a descrição das habilidades que pretende avaliar e, quando deseja um item contextualizado, pode sugerir um possível contexto. Assim, instruções, habilidades e contexto constituem um primeiro *input* para que a IA elabore um item com as características desejadas.

O modelo de conversação adotado pode ser visualizado na Figura 1, considerando que o estudo se limitou a testar a elaboração de itens de múltipla escolha. Em resumo, após a primeira entrada, espera-se que o chat tenha sido instruído a gerar conteúdo com certos padrões e de maneira sequencial, mas cabe ao professor/elaborador revisar e sugerir melhorias à IA quantas vezes for necessário, visando como *output* final um item inédito e isento de erros de conteúdo ou de elaboração. Algumas instruções (comandos) empregados para elaboração de itens discutidos nesse trabalho estão disponíveis em (Silva, 2023).

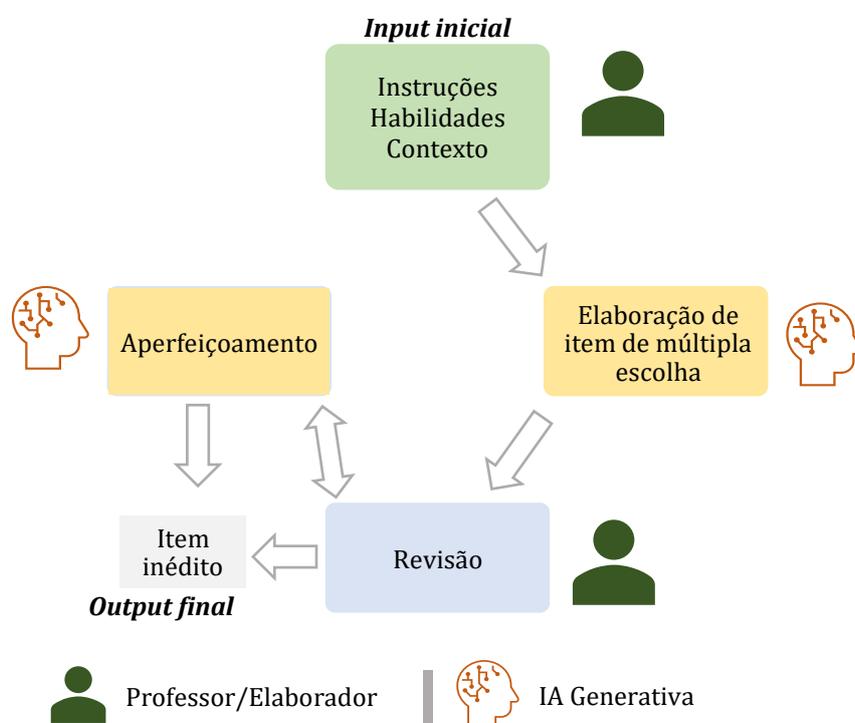


Figura 1 – Processo e agentes envolvidos na elaboração de itens de múltipla escolha com auxílio de IA generativa

Com o objetivo de avaliar alguns itens elaborados com auxílio de IA à luz da Teoria Clássica dos Testes (TCT), aplicou-se um teste de 20 itens de múltipla escolha de matemática. Destes, 9 foram elaborados por terceiros, como o CAEd e Inep, e 11 itens foram gerados com auxílio do Bing Chat e do ChatGPT. Os itens que compuseram o teste atendem algumas das habilidades descritoras preconizadas na Matriz de Referência de Matemática do Sistema de Avaliação da Educação Básica (Saeb), proposta em 2001. Além de fazer parte do estudo, o teste foi essencial para a promoção de intervenções pedagógicas visando o desenvolvimento de habilidades matemáticas importantes nessa

etapa de ensino, como a resolução de problemas envolvendo as operações com números racionais e a modelagem de problemas por meio da linguagem algébrica.

O teste foi respondido por 61 estudantes do 9º ano do ensino fundamental de uma escola pública localizada em Terra Santa, um município no interior do estado do Pará, em outubro de 2023, com duração de 2 tempos de aula de 45 minutos. O professor responsável pelas turmas, que também conduziu a pesquisa, assegurou aos estudantes no momento da aplicação do teste que suas respostas seriam mantidas em estrito anonimato no estudo, garantindo que nenhum dos participantes seria identificado, preservando assim a sua privacidade. A aplicação se deu por meio de um caderno de questões, composto por uma folha de resposta (gabarito) que foi submetido à leitura óptica no aplicativo “Prova Fácil”, o qual gerou uma planilha em formato CSV com as respostas. A análise via TCT das respostas foi feita com auxílio do software RStudio e do pacote *Latent Trait Models under IRT* (ltm).

A análise considerou a proporção de respostas corretas para cada item, utilizada como indicador de dificuldade, o índice clássico de discriminação e a medida de correlação ponto bisserial. O alfa de Cronbach, que é um indicador de consistência interna dos itens e do teste, também foi empregado (Pasquali, 2017). Ademais, utilizou-se a Análise Gráfica de Item (AGI), uma ferramenta que permite avaliar atributos estatísticos de um item de maneira visual (Batenburg; Laros, 2002; Laros, 2017).

4. Resultados

4.1 Gerando itens a partir de um comando básico

Nesta subseção, são apresentados exemplos de itens gerados a partir de comandos básicos, juntamente com algumas de suas limitações. O quadro a seguir exhibe itens destinados a avaliar habilidades especificadas na Base Nacional Comum Curricular (BNCC) para a componente de Matemática do 9º ano. Dois itens foram gerados utilizando as ferramentas ChatGPT e Bing Chat, respectivamente, com base em um comando que estabelece o processo e as informações essenciais para a elaboração do item. Outros comandos detalhados para orientar a criação de itens podem ser encontrados no trabalho de Silva (2023).

A análise do quadro revela que ambas as ferramentas geram itens coerentes com a habilidade proposta, aderindo à estrutura estabelecida, que contém enunciado, comando para resposta e quatro alternativas distintas e compatíveis com o comando. No item gerado pelo ChatGPT, apesar de não apresentar erros de conteúdo e propor um problema envolvente, houve uma falha na apresentação das alternativas de resposta, comprometendo a qualidade do item e resultando em sua exclusão de qualquer certame. Esse é um exemplo de limitação comum no processo de elaboração de item com auxílio deste recurso.

O item gerado pelo Bing Chat se deu a partir de um contexto retirado da internet. Dado o contexto proposto e a habilidade informada, o Bing gerou um item coerente, seguindo rigorosamente o processo estabelecido e gerando quatro alternativas, com

apenas uma correta. No entanto, o gabarito foi indicado incorretamente. Embora o item atenda à habilidade proposta, ele não exige do respondente um processo mental mais elaborado, como a criação de um modelo que emprega o conceito de função. Este item conduz o respondente ao procedimento imediato de cálculo, sendo, portanto, um item relativamente simples que não requer necessariamente o uso do conceito de função.

Comando
<p>Neste chat, a tarefa é criar itens inéditos de múltipla escolha para matemática. Siga os passos abaixo: 1º Passo: Solicite ao usuário que forneça uma descrição da habilidade que o item deve avaliar. 2º Passo: Após receber a habilidade do usuário, peça a ele que forneça um contexto ou cenário que será a base para o item. 3º Passo: Com base nas informações fornecidas, crie um item de múltipla escolha com quatro alternativas, sendo apenas uma correta. O item deve avaliar de forma precisa a habilidade informada e não ser óbvio. Além disso, o enunciado do item deve ser conciso e esclarecer para o respondente o contexto. 4º Passo: Forneça a solução ou gabarito do item de forma detalhada.</p>
Exemplo de item gerado pelo ChatGPT
<p>ChatGPT Habilidade: (EF09MA09) Compreender os processos de fatoração de expressões algébricas, com base em suas relações com os produtos notáveis, para resolver e elaborar problemas que possam ser representados por equações polinomiais do 2º grau. Contexto proposto pelo professor/elaborador: Cercar uma área retangular máxima destinada à produção de hortaliças em uma determinada escola, utilizando 20 metros de tela para delimitação. Enunciado: Um agricultor quer cercar uma área retangular para cultivar hortaliças em uma escola. Ele tem 20 metros de tela para delimitação. Qual é a expressão algébrica que representa a área máxima que ele pode cercar? Alternativas: (A) $A = 5x(20 - x)$ (B) $A = 10x(10 - x)$ (C) $A = 2x(10 - x)$ (D) $A = 4x(20 - x)$ Proposta de resolução do ChatGPT Para resolver esse problema, precisamos maximizar a área do retângulo. Se x representa a largura do retângulo, então o comprimento do retângulo será $10 - x$ (pois a soma da largura e do comprimento deve ser igual a 10 metros, já que ele só tem 20 metros de tela). A área do retângulo é dada por $A = x(10 - x) = 10x - x^2$. Para encontrar a expressão que maximiza a área, devemos derivar a função da área em relação a x e encontrar o valor de x que torna a derivada igual a zero. No entanto, para este item, vamos focar na escolha correta. A alternativa correta é (B) $A = 10x(10 - x)$. Observação: Ausência de gabarito entre as alternativas.</p>
Exemplo de item gerado pelo Bing Chat
<p>Habilidade: (EF09MA06) Compreender as funções como relações de dependência unívoca entre duas variáveis e suas representações numérica, algébrica e gráfica e utilizar esse conceito para analisar situações que envolvam relações funcionais entre duas variáveis. Contexto proposto pelo professor/elaborador: Na missão, são US\$ 58,50 fixos para o vídeo selecionado mais US\$ 0,39 para cada 1 mil impressões (visualizações) pagas pela marca, ou seja, a marca impulsiona os vídeos selecionados e depois te paga por isso. Texto retirado de: <https://g1.globo.com/empreendedorismo/noticia/2023/07/03/as-quatro-maneyras-de-ganhar-dinheiro-no-tiktok.ghtml>. Enunciado: Um influenciador digital recebe US\$ 58,50 fixos para cada vídeo selecionado em uma missão, além de US\$ 0,39 para cada 1 mil impressões pagas pela marca. Se a marca impulsionou um vídeo que teve 5 mil impressões, qual será o valor total que o influenciador receberá? Alternativas: (A) US\$ 60,45 (B) US\$ 61,45 (C) US\$ 62,45 (D) US\$ 63,45 Gabarito: A resposta correta é a alternativa (C). O influenciador receberá US\$ 58,50 fixos pelo vídeo mais US\$ 0,39 para cada 1 mil impressões. Como o vídeo teve 5 mil impressões, ele receberá um adicional de $US\\$ 0,39 * 5 = US\\$ 1,95$. Portanto, o valor total que o influenciador receberá será de $US\\$ 58,50 + US\\$ 1,95 = US\\$ 60,45$. Observação: item muito simples, que não requerem modelagem ou mesmo uso da teoria de funções.</p>

Quadro 1 – Exemplo de comando e itens gerados no ChatGPT e Bing Chat

Fonte: Elaborado pelo próprio Autor com auxílio do ChatGPT (<https://curtlink.com/cPwa>) e Chat Bing.

É pertinente discutir que a seleção da habilidade, conforme definida na BNCC, e os verbos nela incorporados que indicam um processo cognitivo, podem ter uma relação importante com o objetivo avaliativo do item gerado. No exemplo do item gerado pelo Bing Chat, que tem como objetivo avaliar o domínio da habilidade (EF09MA06), os verbos associados são “compreender” e “analisar”. Estes verbos, de acordo com a taxonomia de Bloom (Ferraz; Belhot, 2010), abrangem processos cognitivos que variam de básicos a intermediários. Portanto, a escolha da habilidade, que define o processo cognitivo a ser desenvolvido ou avaliado, pode ter influência no comando para resposta gerado e no nível de dificuldade do item.

4.2 Análise de itens gerados com auxílio de IA generativa

Nesta seção, apresenta-se uma análise da qualidade dos itens do teste aplicado à luz da TCT. Medidas de dificuldade do item (proporção de acerto), discriminação clássica, correlação ponto bisserial e alfa de Cronbach são exibidos na Tabela 1.

Tabela 1 – Parâmetros clássicos dos itens que compõe o teste: proporção de acerto (dificuldade), discriminação (clássica e correlação ponto bisserial) e consistência interna (alfa de Cronbach)

Item	Proporção de acerto	Discriminação clássica (D)	Correlação ponto bisserial	alfa de Cronbach (0,80)
1*	0,59	0,50	0,47	0,80
2*	0,80	0,56	0,50	0,80
3*	0,95	0,06	0,03	0,81
4*	0,77	0,69	0,55	0,79
5**	0,56	0,44	0,38	0,80
6**	0,69	0,56	0,49	0,80
7**	0,57	0,44	0,40	0,80
8**	0,41	0,56	0,53	0,79
9*	0,69	0,56	0,49	0,80
10**	0,36	0,62	0,54	0,79
11**	0,52	0,25	0,22	0,81
12*	0,79	0,50	0,47	0,80
13**	0,31	0,75	0,60	0,79
14**	0,61	0,62	0,50	0,80
15*	0,61	0,69	0,53	0,79
16**	0,49	0,81	0,65	0,79
17*	0,46	0,62	0,51	0,79
18*	0,87	0,38	0,45	0,80
19**	0,75	0,38	0,37	0,80
20**	0,67	0,50	0,44	0,80

Fonte: Elaborado pelo próprio autor. *Itens retirados de outras fontes: Avalia Brasil, Inep ou CAED. **Itens criados ou aperfeiçoado com auxílio de IA Generativa.

O coeficiente alfa de Cronbach do teste de 0,8, mostra uma consistência interna aceitável para o instrumento, conforme estabelecido por Pasquali (2017). Em relação ao poder discriminativo observado, somente os itens 3, 11, 18 e 19 apresentaram uma capacidade reduzida de distinguir entre sujeitos que possuíam um alto nível de habilidade e aqueles que provavelmente não demonstraram domínio das habilidades avaliadas – corroborado pelas medidas de discriminação inferiores a 0,4 (Rabelo, 2013). Entre esses itens, os de número 11 e 19 foram elaborados ou aprimorados com o auxílio da Inteligência Artificial.

Um teste de comparação de médias foi realizado utilizando o teste t de Student para avaliar as diferenças nas médias do índice clássico de discriminação entre os itens provenientes de fontes externas ($\mu^*=0,54$) e os itens gerados ou melhorados por meio da Inteligência Artificial ($\mu^{**}=0,50$). Os resultados indicaram que a qualidade dos itens criados com o auxílio da IA, em termos de capacidade discriminativa, não apresenta diferença estatisticamente significativa em comparação com os itens oriundos de fontes externas ($t = 0,40$; $df = 15,64$; $\text{valor-p} = 0,69$). Os índices de correlação ponto bisserial dos itens reforçam que a maioria deles possui um poder discriminativo adequado, com exceção dos itens 3, 5, 11, 18 e 19.

Para exemplificar como os itens objetivos gerados por IA podem ser analisados do ponto de vista pedagógico, apresenta-se dois desses itens: 11 e 13. Esse último (Figura 2) foi gerado com o auxílio do Bing Chat. Ele demonstrou um poder discriminativo satisfatório e um alto nível de dificuldade para a maioria dos participantes, com uma taxa de acerto de apenas 31%.

O gráfico de linhas exibido na Figura 2 é o recuso visual empregado para a análise de itens – a AGI (Laros, 2017). A AGI do item 13, portanto, permite observar uma tendência de crescimento da proporção de respostas ao gabarito (alternativa B) em função do aumento do escore. Significa que quanto mais habilidoso é um grupo de estudantes, maiores são suas chances de resolver o item corretamente, o que é esperado na TCT. Tal constatação sugere que o item em análise é adequado para avaliação da habilidade proposta. Outro indicador exibido na AGI é o de correlação ponto bisserial da alternativa correta (0,52), que assegura o bom poder discriminativo do item.

Do ponto de vista do conteúdo abordado, o item 13 busca avaliar se os educandos são capazes de “(D27) Efetuar cálculos simples com valores aproximados de radicais”. A tarefa gerada pela IA para avaliar essa habilidade foi o cálculo aproximado de $\sqrt{82} - 45$. Embora seja um procedimento relativamente simples, o gráfico de barras, que representa a frequência de respostas por alternativas, mostra que o distrator D mostrou-se atrativo para os participantes com baixa proficiência, sugerindo que uma proporção significativa dos avaliados pode ter realizado o cálculo subtraindo 82-45, sem a operação prévia com o radical. Este item pode indicar uma falta de familiaridade ou compreensão do conceito

de raiz quadrada por parte de um grupo de participantes, o que requer atenção especial do professor.

Item 13 – Descritor: (D27) Efetuar cálculos simples com valores aproximados de radicais.

(D27 - IA/PROFESSOR) Qual é o resultado aproximado de

$$\sqrt{82} - 45 ?$$

(A) -37.

(B) -36.

(C) 36.

(D) 37.

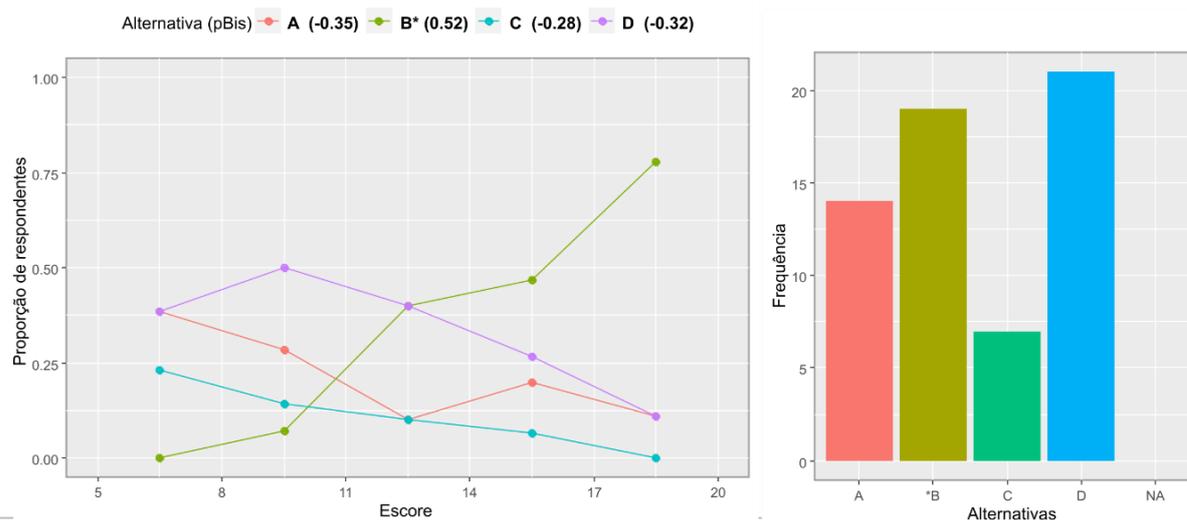


Figura 2 – Item 13 e sua análise gráfica

É interessante notar que nesse processo de elaboração de item o Bing Chat foi capaz de propor uma tarefa compatível com a habilidade informada, além de gerar alternativas satisfatórias para resposta.

O item 11 do teste (Figura 3), também gerado com auxílio de IA, já não apresenta um bom poder discriminativo, indicado por uma baixa medida de correlação ponto bisserial e pelo comportamento das respostas demonstrado na AGI, em que não há uma tendência de crescimento da proporção de respostas ao gabarito (alternativa D) em função do aumento do escore, considerando escores inferiores a 13 acertos. Percebe-se que mesmo indivíduos altamente proficientes podem ter tido dificuldade ao resolver o

problema proposto. Cerca de 52% dos participantes responderam corretamente ao item, o que indica um nível médio de dificuldade.

No que se refere ao conteúdo do item, ele busca aferir a capacidade dos sujeitos avaliados de “(D26) Resolver problema com números racionais envolvendo as operações (adição, subtração, multiplicação, divisão, potenciação)”. O problema proposto atende precisamente a habilidade informada, é ausente de erros conceituais e possui alternativas coerentes com problema. O item 13 permitiu constatar que metade dos estudantes enfrenta obstáculos ao lidar com o problema proposto. Esses obstáculos podem estar relacionados à falta de domínio dos procedimentos matemáticos necessários ou à dificuldade na interpretação do enunciado e das instruções para a resposta.

Item 11 – (D26) Resolver problema com números racionais envolvendo as operações (adição, subtração, multiplicação, divisão, potenciação).

(D26 – IA/PROFESSOR) Uma embarcação tem um tanque de combustível com capacidade para 120 litros. Em uma determinada viagem, a embarcação partiu com tanque completamente abastecido e chegou ao seu destino com $\frac{2}{3}$ de sua capacidade. Quantos litros restaram no tanque ao final da viagem?

- (A) 30.
- (B) 40.
- (C) 60.
- (D) 80.

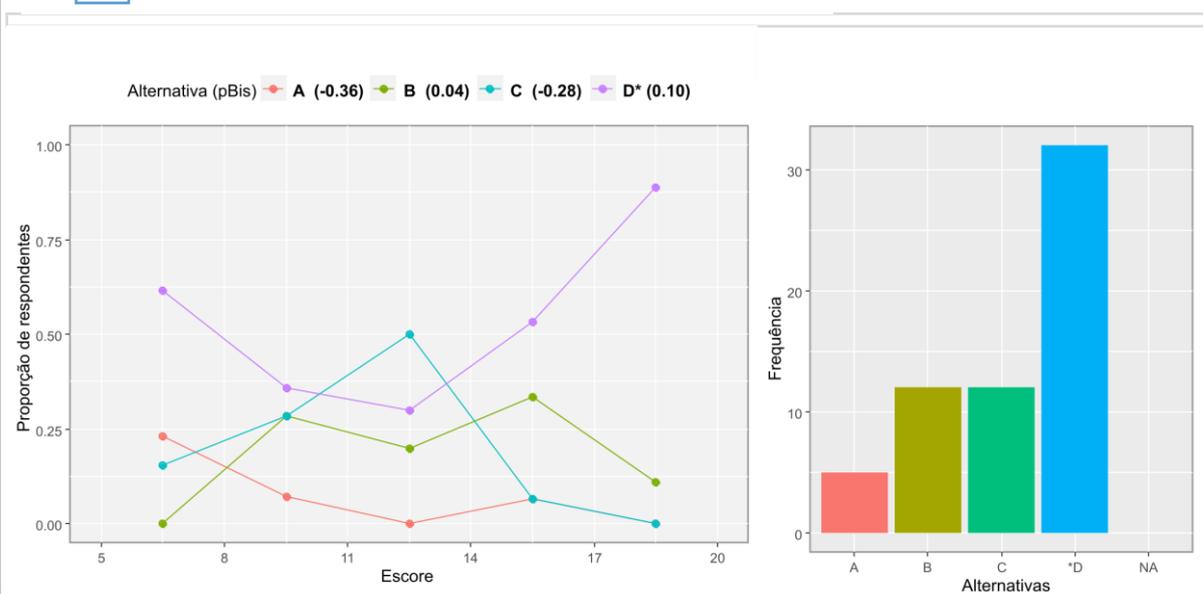


Figura 3 – Item 11 e sua análise gráfica

5. Conclusões

A elaboração de itens é um processo criativo que demanda do elaborador não apenas um conhecimento técnico sobre a produção de itens, mas também uma compreensão sólida do conteúdo e do currículo em questão. Este processo pode ser otimizado com o uso de chats baseados em IA generativa, que possuem a capacidade de gerar textos coerentes semelhantes aos que são produzidos por humano, seguindo orientações fornecidas em forma de comando textual (Silva, 2023).

O modelo básico de elaboração de itens proposto neste artigo atribui ao docente/elaborador o papel de produzir um comando que oriente o processo de produção de itens em chats, realizar a curadoria das habilidades avaliadas e do contexto a ser considerado pelo item, além da avaliação crítica dos itens produzidos. Ou seja, os itens precisam ser revisados cuidadosamente para garantir a conformidade com as diretrizes mencionadas por INEP (2010), CAED (2008) e Vianna (1978), o que requer do elaborador pleno domínio das boas práticas de elaboração e revisão de itens.

Este estudo revelou limitações dos chats, principalmente na redação das alternativas e na produção de um comando para resposta que mobilizam processos cognitivos superiores, como a capacidade de analisar, avaliar e criar dos respondentes (Ferraz; Belhot, 2010). Embora parte dos itens gerados estejam em conformidade com o contexto proposto pelo elaborador, alguns dos comandos para resposta gerados possuem um cunho generalista e simplista.

Ademais, a otimização do processo de elaboração de itens com o auxílio desses recursos, sem envolver um refinamento delas para cumprir precisamente tais tarefas, pode requerer o aperfeiçoamento do comando inicial para elaboração de item para que considere os processos cognitivos previstos na taxonomia de Bloom, que não foi considerada no presente estudo. Esta possibilidade fica como sugestão para trabalhos futuros no campo da educação matemática.

A avaliação de um conjunto de itens gerados, considerando a capacidade discriminativa, o nível de dificuldade e a consistência interna, foi realizada por meio da análise de um teste aplicado a um conjunto de estudantes. Os resultados indicam que os itens são satisfatórios. Uma simples comparação do índice de discriminação entre os itens gerados e os itens produzidos por especialistas em elaboração de itens revela uma qualidade similar. Portanto, tais constatações evidenciam o potencial promissor da abordagem proposta neste estudo para a elaboração de itens com o auxílio de Inteligência Artificial generativa no campo da educação matemática.

6. Referências

- BATENBURG, T. A. van; LAROS, J. A. Graphical Analysis of Test Items. **Educational Research and Evaluation**, v. 8, n. 3, p. 319–333, 2002. Disponível em: <<https://doi.org/10.1076/edre.8.3.319.3856>>. Acesso em: 9 fev. 2024.
- BOULAY, du B. Artificial Intelligence in Education and Ethics. *In*: HANDBOOK OF OPEN, DISTANCE AND DIGITAL EDUCATION. Singapore: **Springer Nature Singapore**, 2022. p. 1–16. Disponível em: <https://doi.org/10.1007/978-981-19-0351-9_6-2>. Acesso em: 10 de jun. 2023.
- BUCHANAN, B. G. A (very) brief history of artificial intelligence. **Ai Magazine**, v. 26, n. 4, p. 53, 2005. Disponível em: <<https://cutt.ly/Qw7a3lNu>>. Acesso em: 25 jun. 2023.
- CAED - Centro de Políticas Públicas e Avaliação da Educação. Universidade Federal de Juiz de Fora. **Guia de Elaboração de Itens: matemática**. 2008. Disponível em: <<https://cutt.ly/bw7a35b4>>. Acesso em: 20 jun. 2023.
- DELIPETREV, B.; TSINARAKII, C.; KOSTIĆ, U. **Historical Evolution of Artificial Intelligence**, 2020. Disponível em: <<https://eprints.ugd.edu.mk/28050>>. Acesso em: 12 de jul. 2023.
- FERRAZ, A. P. do C. M.; BELHOT, R. V. Taxonomia de Bloom: revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais. **Gestão & Produção**, v. 17, n. 2, p. 421–431, 2010. Disponível em: <<https://doi.org/10.1590/S0104-530X2010000200015>>. Acesso em: 10 de fev. de 2024.
- FIGUEIREDO, L. de O. *et al.* Desafios e impactos do uso da Inteligência Artificial na educação. **Educação Online**, v. 18, n. 44, p. e18234408, 2023. Disponível em: <<https://doi.org/10.36556/eol.v18i44.1506>>. Acesso em: 25 de mar. 2024.
- GATTI, F. N. **Educação básica e inteligência artificial: perspectivas, contribuições e desafios**. 2019. Dissertação (Mestrado em Educação: currículo) - Pontifícia Universidade Católica de São Paulo, São Paulo, 2019. Disponível em: <<https://tede2.pucsp.br/handle/handle/22788>>. Acesso em: 24 jun. 2023.
- GIMENES, N. *et al.* Além da Prova Brasil: investimento em sistemas próprios de avaliação externa. **Estudos em Avaliação Educacional**, v. 24, n. 55, p. 12, 2013. Disponível em: <<http://publicacoes.fcc.org.br/ojs/index.php/eae/article/view/2718>>. Acesso em: 12 de fev. 2023.
- GIRAFFA, L.; KHOLS-SANTOS, P. Inteligência Artificial e Educação: conceitos, aplicações e implicações no fazer docente. **Educação em Análise**, v. 8, n. 1, p. 116–134, 2023. Disponível em: <<https://doi.org/10.5433/1984-7939.2023v8n1p116>>. Acesso em: 20 de fev. 2024.
- HAENLEIN, M.; KAPLAN, A. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. **California Management Review**, v. 61, n. 4, p. 5–14, 2019. Disponível em: <<https://doi.org/10.1177/0008125619864925>>. Acesso em: 3 jun. 2023.

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Ministério da Educação. 2010. **Guia de Revisão e Elaboração de Itens**. Disponível em: <<https://cutt.ly/Vw7a2518>>. Acesso em: 24 mai. 2023.

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Ministério da Educação. 2018. **Saeb: documentos de referência -versão 1.0**. Disponível em: <<http://portal.inep.gov.br/web/guest/educacao-basica/saeb>> Acesso em: 10 de jun. 2023.

LAROS, J. A. Análise Gráfica de Itens. *In: Psicometria: Teoria dos Testes na Psicologia e na Educação*. Petrópolis: Vozes, 2017.

LO, C. K. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. **Education Sciences**, v. 13, n. 4, p. 410, 2023. Disponível em: <<https://doi.org/10.3390/educsci13040410>>. Acesso em: 2 de fev. 2024.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. 3. ed. Rio de Janeiro: Elsevier, 2013.

SAMOILI, S. *et al.* **Defining Artificial Intelligence 2.0: Towards an operational definition and taxonomy for the AI landscape**. 2021. Disponível em: <<https://doi.org/10.2760/019901>>. Acesso em: 1 jul. 2023.

SANTOS, R. P.; SANT'ANA, C. de C.; SANT'ANA, I. P. O ChatGPT como recurso de apoio no ensino da Matemática. **Revemop**, v. 5, p. e202303, 2023. Disponível em: <<https://doi.org/10.33532/revemop.e202303>>. Acesso em: 10 de jan. 2024.

SEJNOWSKI, T. J. Large Language Models and the Reverse Turing Test. **Neural Computation**, v. 35, n. 3, p. 309–342, 2023. Disponível em: <https://doi.org/10.1162/neco_a_01563>. Acesso em: 9 jul. 2023.

SILVA, A. C. da. **Desafios e possibilidades do uso de inteligência artificial generativa na elaboração e revisão de itens de matemática**. 2023. Dissertação (Mestrado em Matemática) - Universidade Federal do Oeste do Pará, Santarém, 2023. Disponível em: <<http://bit.ly/3xAo8CU>>. Acesso em: 9 fev. 2024.

SOUSA, C. P. de. Dimensões da avaliação educacional. **Estudos em Avaliação Educacional**, n. 22, p. 101, 2000. Disponível em: <<https://doi.org/10.18222/eae02220002218>>. Acesso em: 9 out. 2023.

TELES, L.; NAGUMO, E. Uma inteligência artificial na educação para além do modelo behaviorista. **Revista Ponto de Vista**, v. 12, n. 3, p. 01–15, 2023. Disponível em: <<https://doi.org/10.47328/rpv.v12i3.15452>>. Acesso em: 3 de fev. 2024.

YU, H. Reflection on whether Chat GPT should be banned by academia from the perspective of educativo and teaching. **Frontiers in Psychology**, v. 14, 2023. Disponível em: <<https://doi.org/10.3389/fpsyg.2023.1181712>>. Acesso em: 30 jun. 2023.