Revista Intersaberes ISSN: 1809-7286
Vol. 21 e26do302 2026

Avaliação da Qualidade de Itens de Matemática Gerados por Inteligência Artificial Generativa: um estudo com base na teoria clássica dos testes

Quality Assessment of Mathematics Items Generated by Generative artificial intelligence: a study based on classical test theory

Evaluación de la calidad de ítems matemáticos generados por inteligencia artificial generativa: un estudio basado en la teoría clásica de tests

Fernando Éder Andrade de Lima
Licenciado em Matemática pela Universidade Estadual do Ceará (UECE), mestrando em Ensino de Ciências e Matemática (PPGECM/IFCE), Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
https://orcid.org/0009-0001-2585-098X ; E-mail: fernando.eder01@aluno.ifce.edu.br

Juscileide Braga de Castro
Doutora em Educação (UFC), professora Adjunta da Universidade Federal do Ceará (UFC), na Faculdade de Educação, professora no Programa de Pós-Graduação em Ensino de Ciências e Matemática (PPGECM/IFCE), Universidade Federal do Ceará (UFC)
https://orcid.org/0000-0002-6530-4860 ; E-mail: juscileide@virtual.ufc.br

RESUMO

Este estudo tem como objeto a elaboração e a avaliação de itens de Matemática gerados por Inteligência Artificial Generativa. O objetivo é analisar a qualidade técnica, cognitiva e psicométrica desses itens, à luz de referenciais da Educação Matemática, da Avaliação Educacional e da Psicometria. Trata-se de uma pesquisa exploratória, desenvolvida em ambiente computacional controlado, com base em dados simulados. Quatro sistemas de IA (ChatGPT, Gemini, Perplexity e DeepSeek) foram utilizados para gerar itens de múltipla escolha alinhados à habilidade EF09MA09 da BNCC, analisados segundo a Taxonomia de Bloom Revisada e a Teoria Clássica dos Testes. Os resultados indicam coerência técnica dos itens, embora revelem variações no nível cognitivo mobilizado, na qualidade dos distratores e no desempenho psicométrico estimado, evidenciando a necessidade de mediação docente e de protocolos metodológicos rigorosos.

Palavras-chave: inteligência artificial generativa; educação matemática; avaliação educacional; teoria clássica dos testes.

ABSTRACT

This study examines the development and evaluation of Mathematics items generated by Artificial Intelligence. The objective is to analyze the technical, cognitive, and psychometric quality of these items based on frameworks from Mathematics Education, Educational Assessment, and Psychometrics. This exploratory research was conducted in a controlled computational environment using simulated data. Four AI systems (ChatGPT, Gemini, Perplexity, and DeepSeek) were employed to generate multiple-choice items aligned with the BNCC skill EF09MA09 and analyzed according to the Revised Bloom's Taxonomy and Classical Test Theory. The results indicate technical coherence of the items, while revealing variations in cognitive demand, distractor quality, and estimated psychometric performance, highlighting the need for teacher mediation and rigorous methodological protocols.

Keywords: generative artificial intelligence; mathematics education; educational assessment; classical test theory.

RESUMEN

Este estudio tiene como objeto la elaboración y evaluación de ítems de Matemáticas generados por Inteligencia Artificial. El objetivo es analizar la calidad técnica, cognitiva y psicométrica de dichos ítems, a partir de referentes de la Educación Matemática, la Evaluación Educativa y la Psicometría. Se trata de una

investigación exploratoria, desarrollada en un entorno computacional controlado, basada en datos simulados. Cuatro sistemas de IA (ChatGPT, Gemini, Perplexity y DeepSeek) fueron utilizados para generar ítems de opción múltiple alineados con la habilidad EF09MA09 de la BNCC, analizados según la Taxonomía de Bloom Revisada y la Teoría Clásica de los Tests. Los resultados indican coherencia técnica, aunque evidencian variaciones en el nivel cognitivo, la calidad de los distractores y el desempeño psicométrico estimado, reforzando la necesidad de mediación docente y protocolos metodológicos rigurosos.

Palabras-clave: inteligencia artificial generativa; educación matemática; evaluación educativa; teoría clásica de los tests.

INTRODUÇÃO

A inserção das Tecnologias Digitais (TD) no contexto educacional tem se intensificado nas últimas décadas, especialmente no apoio ao planejamento didático, à produção de materiais pedagógicos e à elaboração de atividades avaliativas. Observa-se ainda o uso das TD para a representação de conceitos e simulação, especialmente por meio de *softwares* dinâmicos e Recursos Educacionais Digitais, configurando-se como elementos que podem contribuir nas práticas de ensino e aprendizagem da Matemática (Castro-Filho; Freire; Castro, 2017; Pereira; Scherer, 2022). Recentemente, observa-se a ampliação do cenário das TD com a incorporação da Inteligência Artificial (IA) ao contexto educacional, a qual vem oferecendo suporte automatizado e adaptativo para o planejamento didático, a criação de recursos pedagógicos e o desenvolvimento de atividades avaliativas (Ribeiro; Navarro; Kalinke, 2024; Silva; Sant'ana; Sant'ana, 2024; Silva; Kampff, 2023).

Há uma mudança qualitativa no uso das tecnologias, deslocando o foco de ferramentas meramente operacionais para sistemas capazes de apoiar processos criativos e autorais no trabalho docente (Aguirre, 2025; Castro, 2024; Silva; Kampff, 2023; Santos; Limoni; Souza, 2024). No cotidiano escolar, as IAs têm sido exploradas como ferramentas de apoio, auxiliando docentes na criação de exercícios e materiais pedagógicos, prática já adotada por 76% dos professores brasileiros (Fundação Itaú Equidade.Info, 2024). Esses recursos permitem adaptar níveis de complexidade e a personalização das atividades, ajustando o conteúdo ao ritmo e às necessidades individuais de aprendizagem dos estudantes (CIEB, 2024).

Estudos indicam que professores tendem a perceber de forma favorável o uso de sistemas baseados em Inteligência Artificial, especialmente quando contribuem para a otimização do tempo pedagógico e para a personalização das atividades escolares (Santos;

Limoni; Souza, 2024; Silva; Kampff, 2023). Santos, Limoni e Souza (2024) corroboram essa percepção ao destacarem que a aceitação da IA decorre de sua capacidade de agilizar o acesso à informação, apoiar a gestão das rotinas docentes e subsidiar processos de organização curricular e avaliação diagnóstica. Convergentemente, Silva e Kampff (2023) argumentam que a IA pode ser compreendida como um recurso de apoio à docência, ao simplificar a produção de materiais educacionais e reduzir a carga operacional do professor. Segundo esses autores, essa economia de tempo cria condições para o docente concentrar seus esforços na análise crítica dos recursos utilizados e na mediação de aprendizagens no processo educativo.

Contudo, a incorporação dessas ferramentas não ocorre de maneira isenta de tensões, especialmente no que se refere à confiabilidade dos produtos gerados, à presença de vieses algorítmicos, às implicações éticas e à necessidade de formação técnico-pedagógica adequada para seu uso responsável (Ribeiro; Navarro; Kalinke, 2024; Aguirre, 2025; Silva Junior; Quartieri, 2025; Vasconcelos *et. al.*, 2025). Nesse sentido, essas preocupações tornam-se relevantes ao se considerar o uso da IA na elaboração de itens avaliativos de Matemática. A construção de itens de múltipla escolha de Matemática exige rigor conceitual, clareza linguística, adequação cognitiva e coerência didática (CAEd, 2008), aspectos que nem sempre são plenamente atendidos quando ferramentas de IA são utilizadas sem critérios explícitos ou referenciais consolidados.

Embora as IAs apoiem a elaboração de itens matemáticos, ainda carecem de ancoragem sistemática em referenciais consolidados, como a Taxonomia de Bloom Revisada e a Teoria Clássica dos Testes (TCT), tanto na formulação dos *prompts* quanto na análise da qualidade técnica e cognitiva das questões (Silva; Tanaka Filho, 2025). A Taxonomia de Bloom Revisada (Bloom *et. al.*, 1956; Anderson; Krathwohl, 2001) organiza os processos cognitivos em seis níveis hierarquizados — lembrar, compreender, aplicar, analisar, avaliar e criar — permitindo examinar a complexidade demandada pelo item. Já a TCT fornece indicadores psicométricos, como dificuldade e discriminação, que possibilitam avaliar o desempenho técnico das questões.

Além disso, observa-se uma lacuna específica na literatura no que se refere ao papel atribuído às IAs nos processos avaliativos. Em geral, essas ferramentas são empregadas apenas como geradoras de itens, sem que se explore de forma sistemática sua

potencialidade para atuar também como avaliadoras técnicas e cognitivas dos próprios produtos que criam.

Nesse sentido, pesquisas apontam para a escassez de estudos sobre o uso de tecnologias como suporte à avaliação técnica, destacando uma lacuna na eficácia dos métodos avaliativos (Paes *et. al.*, 2025) e enfatizam a necessidade, especialmente para suprir métodos convencionais, nos quais os docentes negligenciam, frequentemente, o rigor e a qualidade psicométrica das questões (Paes *et. al.*, 2025; Soares; Emiliano; Soares, 2020).

A investigação dessa possibilidade, quando fundamentada em referenciais normativos como o Guia de Elaboração de Itens do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e métricas psicométricas, pode contribuir para o desenvolvimento de protocolos metodológicos mais rigorosos de elaboração e pré-validação de instrumentos avaliativos. Diante desse contexto, este estudo investiga em que medida itens de Matemática autogerados por diferentes IAs apresentam qualidade técnica, coerência cognitiva e comportamento psicométrico relevante, quando analisados em ambiente computacional controlado, à luz da TCT.

O objetivo do artigo consiste em analisar a qualidade técnica, cognitiva e psicométrica desses itens, à luz de referenciais da Educação Matemática, da Avaliação Educacional e da Psicometria. Para isso, o estudo propõe a elaboração de dois *prompts* estruturados para geração e avaliação dos itens, bem como a realização de simulações computacionais de respostas hipotéticas para estimativa de indicadores psicométricos.

O artigo está organizado da seguinte forma: a seguir, apresenta-se o referencial teórico, que discute o uso da IA no contexto educacional, com ênfase na elaboração de itens. Em seguida, descreve-se a metodologia do estudo, detalhando o delineamento do experimento e os procedimentos adotados. Na sequência, são apresentados e discutidos os resultados obtidos. Por fim, as considerações finais sintetizam os principais achados e indicam possibilidades para estudos futuros.

INTELIGÊNCIA ARTIFICIAL EM PRÁTICAS AVALIATIVAS EM MATEMÁTICA

A integração da IA no campo educacional é discutida na literatura como um recurso de apoio ao planejamento pedagógico, à produção de materiais didáticos e, mais recentemente, à elaboração de itens avaliativos, inserindo-se na necessidade da escola

moldar-se diante das mudanças tecnológicas (Bonservizi; Sgreccia, 2021; Teixeira *et. a*l., 2022). Nesse cenário, estudos indicam que ferramentas como *ChatGPT, Gemini* e *Perplexity* podem auxiliar docentes na criação de atividades diversificadas, desde que seu uso seja orientado por critérios pedagógicos explícitos e mediado por avaliação humana qualificada (Silva; Tanaka Filho, 2025; Pedrini, 2025).

Pesquisas recentes sugerem que itens de Matemática gerados por IA podem alcançar níveis satisfatórios de qualidade técnica e poder discriminativo, especialmente quando sua elaboração é orientada por *prompts* estruturados e alinhados a referenciais consolidados, como a Taxonomia de Bloom Revisada e a TCT (Silva; Tanaka Filho, 2025). Esses achados apontam para o potencial da IA como instrumento complementar ao trabalho docente, sem que isso implique sua substituição em decisões finais dos processos avaliativos.

Entretanto, a literatura destaca que a simples adoção dessas tecnologias não garante a qualidade pedagógica dos produtos gerados (Lima Netto, 2024). A ausência de critérios técnicos e pedagógicos claros na elaboração de comandos pode resultar em materiais com falhas conceituais, ambiguidades linguísticas ou desalinhamento entre a habilidade avaliada e o nível cognitivo requerido do estudante (Silva; Tanaka Filho, 2025; Del Valle, 2025). Tais riscos são criados pelo fenômeno das "alucinações artificiais", em que o sistema pode gerar informações incorretas ou gabaritos imprecisos, tornando a mediação humana e a curadoria docente indispensáveis para assegurar a integridade do ensino (Silva, Sant'Ana; Sant'Ana, 2024; Costa; Moraes, 2024). Há de se assinalar ainda os aspectos culturais e sociais que não costumam ser considerados pelas IAs, requisitando, portanto, uma atenção maior do professor para evitar contextos universalistas ou vieses que não condizem com a realidade (Aguirre, 2025; Ribeiro; Navarro; Kalinke, 2024; Silva; Sá, 2024).

A eficácia do uso da IA no contexto educacional está diretamente associada ao desenvolvimento do letramento em IA (Vasconcelos *et. al.,* 2025) e ao domínio de estratégias de engenharia de *prompt* (*engineer prompt*), competências essenciais para o docente formular instruções claras e pedagogicamente orientadas (Zhan *et. al.*, 2025; Silva; Tanaka Filho, 2025). A ausência de formação específica que contemple domínio tecnológico contribui para o uso dessas tecnologias a uma abordagem meramente

instrumental, favorecendo a dependência tecnológica e a diminuição da reflexão crítica de professores e estudantes (Pedrini, 2025; Belkina *et. al.*, 2025; Bido; Wiese; Nakamura, 2024).

Nesse contexto, estudos indicam que a IA deve ser compreendida como um recurso complementar de apoio ao trabalho docente, cuja utilização requer supervisão humana constante para minimizar vieses e limitações algorítmicas (Silva; Sant'Ana; Sant'Ana, 2024; Santos; Limoni; Souza, 2024). Assim, a formação de professores de Matemática deve assumir uma perspectiva crítica e reflexiva, preparando o educador para atuar como designer de experiências de aprendizagem, capaz de integrar a tecnologia intencionalmente, ampliando a autonomia e o raciocínio lógico dos estudantes, sem comprometer o protagonismo docente frente à automação (Kehoe, 2023; Matos; Coutinho, 2024; Belkina *et. al.*, 2025).

METODOLOGIA

Este estudo caracteriza-se como uma pesquisa exploratória, cujo objetivo é proporcionar maior familiaridade com o problema investigado, ampliar a compreensão do fenômeno e subsidiar estudos posteriores. Gil (2008) descreve a pesquisa exploratória como um procedimento flexível voltado ao desenvolvimento, esclarecimento e refinamento de conceitos e ideias para formulação de problemas e hipóteses futuras.

A utilização de dados simulados decorre do caráter exploratório da pesquisa e visa testar, por meio de protocolo estruturado, a viabilidade da análise técnica e psicométrica de itens antes de sua aplicação empírica. A simulação considerou níveis diferenciados de proficiência (baixa, média e alta), probabilidades graduais de acerto, distratores plausíveis e variabilidade aleatória, buscando reproduzir padrões típicos de contextos avaliativos reais.

O protocolo metodológico foi organizado em três etapas: (1) elaboração de itens por IA com prompts estruturados; (2) simulação de respostas e estimativas psicométricas com base na TCT; e (3) avaliação cruzada (cross-evaluation) entre diferentes IAs. Essa estrutura assegura transparência, sistematicidade e potencial de replicação.

A análise integrou dimensões técnicas e psicométricas, incluindo exame individual dos itens e avaliação cruzada, permitindo verificar tanto a qualidade das questões quanto a capacidade crítica das IAs sobre suas próprias produções. A amostra de cinco itens

possibilitou aprofundamento analítico, considerando que a geração automatizada requer revisão criteriosa para identificação de falhas técnicas e conceituais (Laverghetta Jr.; Licato, 2023; Mead; Zhou, 2024).

A adoção de indicadores psicométricos da TCT justificou-se como estratégia para estimar o comportamento esperado dos itens e verificar se apresentam propriedades equivalentes ou superiores às de itens elaborados por humanos (Laverghetta Jr.; Licato, 2023; Wróblewska *et. al*., 2025).

- O Índice de Dificuldade (P) corresponde à proporção de respondentes que acertam um determinado item, assumindo valores no intervalo entre 0 e 1. Valores elevados indicam itens mais fáceis, enquanto valores reduzidos indicam maior complexidade (Pasquali, 2017). Para fins de interpretação, adotaram-se os seguintes critérios: Proporção de acertos (0-1). Critérios: P > 0,70 = fácil; 0,30-0,70 = média; P < 0,30 = difícil (Condé, 2001).
- O Índice de Discriminação (D) expressa a capacidade do item de diferenciar respondentes com distintos níveis de proficiência, sendo tradicionalmente estimado a partir da diferença de desempenho entre grupos de alto e baixo escore total (Pasquali, 2017). Neste estudo, valores de diferença de desempenho entre grupos de alto e baixo escore total. Critérios: D ≥ 0,40 = bom; D: 20-0,39 = moderado; D < 0,20 = insuficiente (Condé, 2001).
- A Correlação Ponto-Bisserial (r_pb) mede a relação entre o acerto no item (variável dicotômica) e o escore total no teste (variável contínua), constituindo um importante indicador de consistência interna (Pasquali, 2017). Correlação igual ou superior no item e escore total do teste. Critérios: r_pb ≥ 0,30 = adequado; 0,20-0,29 = marginal; r_pb < 0,20 = problemático (Condé, 2001).

Cabe ressaltar que, por se tratar de um estudo exploratório com dados simulados, os índices apresentados são apenas estimativas para avaliar a coerência técnica, o potencial discriminativo e a plausibilidade estatística dos itens produzidos pelas IAs.

Sistemas de IA Analisados

O experimento selecionou quatro sistemas de IA: *ChatGPT*, *Gemini*, *Perplexity* e

DeepSeek, em razão da ampla difusão e consolidação em contextos educacionais e acadêmicos, o que os torna representativos das soluções atualmente disponíveis para docentes e pesquisadores. Cada sistema foi tratado como uma unidade de análise independente, permitindo comparar tanto a elaboração dos itens quanto os procedimentos adotados na análise cognitiva, técnica e psicométrica. As características técnicas dos sistemas analisados, incluindo empresa desenvolvedora, versões e parâmetros declarados, encontram-se sintetizadas no Quadro 1.

Quadro 1: características técnicas das IAs selecionadas.

Modelo de IA	Parâmetros	Versão	Link de Acesso
ChatGPT	~1.76T (estimado GPT-4)	GPT-4o (maio/2024)	chat.openai.com
Gemini	Não divulgado (~2T est.)	Gemini 1.5 Pro/Flash	gemini.google.com
Perplexity	70B (pplx-70b-online)	pplx-70b-online (2023)	perplexity.ai
DeepSeek	236B total / 21B ativo	DeepSeek-V2 (2024)	deepseek.com

Fonte: Elaborado pelos autores (2026)

Ressalta-se que tais informações do Quadro 1 baseiam-se em dados publicamente divulgados pelas empresas responsáveis, não sendo objeto de validação experimental nem de escolha destas ferramentas neste estudo. A utilização de múltiplas IAs possibilitou observar convergências e divergências nos resultados em um mesmo contexto experimental, sem a interferência de variáveis externas associadas a aplicações educacionais reais.

Procedimentos de Elaboração e Avaliação dos Itens

A etapa de elaboração e avaliação dos itens foi orientada por dois conjuntos de instruções (*prompts*), construídos a partir de referenciais consolidados da Educação Matemática, da Avaliação Educacional e da Psicometria. No campo da Educação Matemática, consideram-se princípios relacionados à definição de habilidades e objetos de conhecimento, bem como à coerência conceitual dos itens e à plausibilidade pedagógica dos distratores, conforme discutido em estudos sobre análise qualitativa de itens e aprendizagem matemática (Soares; Emiliano; Soares, 2020).

No âmbito da Avaliação Educacional, os *prompts* incorporaram critérios técnicos amplamente utilizados em avaliações internas e externas, como clareza do enunciado, alinhamento entre comando e habilidade avaliada, unicidade do gabarito e adequação do nível de complexidade cognitiva, aspectos recorrentes na literatura sobre qualidade de itens e uso pedagógico dos resultados avaliativos (Delmiro; Menezes; Borges Neto, 2024). Já do ponto de vista psicométrico, a elaboração dos *prompts* foi fundamentada nos pressupostos da TCT, especialmente no que se refere à análise do índice de dificuldade, do poder discriminativo dos itens e da consistência interna do instrumento (Condé, 2001).

Na primeira etapa, cada IA foi responsável pela elaboração de um item de Matemática de múltipla escolha sobre o tema de produto notável, cuja habilidade de referência da Base Nacional Comum Curricular (BNCC) é a EF09MA05, totalizando 4 itens gerados artificialmente, orientada pelo *prompt 1* que estabelecia critérios técnicos e pedagógicos explícitos, conforme se observa no *Prompt* 1.

Prompt 1 – Construção do Item de Matemática

Você atuará como especialista em Educação Matemática e Avaliação Educacional, com domínio da Taxonomia de Bloom Revisada e das diretrizes para elaboração de itens avaliativos. Sua tarefa é elaborar um item de Matemática de múltipla escolha, tecnicamente adequado para avaliações diagnósticas destinadas a estudantes do 9° ano do Ensino Fundamental.

Diretrizes obrigatórias para a construção do item

1. Estrutura do item

Enunciado claro, objetivo e contextualizado, adequado à faixa etária do 9° ano do Ensino Fundamental;

Comando explícito que indique com precisão a ação cognitiva esperada;

Quatro alternativas (A, B, C, D);

Apenas uma alternativa correta;

Distratores plausíveis, construídos a partir de erros conceituais ou procedimentais recorrentes em conteúdos algébricos, tais como: equívocos na identificação de fatores, aplicação incorreta de produtos notáveis; falhas na manipulação algébrica de expressões.

2. Referenciais pedagógicos

O item deve estar explicitamente alinhado à habilidade EF09MA09 da BNCC, que trata da compreensão dos processos de fatoração de expressões algébricas e de sua relação com produtos notáveis;

Indique o nível cognitivo predominante mobilizado, conforme a Taxonomia de Bloom Revisada;

O comando do item deve mobilizar claramente esse processo cognitivo, evitando a mera aplicação mecânica de procedimentos.

3. Critérios técnicos

Evite ambiguidades linguísticas no enunciado e nas alternativas;

Evite pistas que indiquem a alternativa correta;

Garanta a unicidade do gabarito;

As alternativas devem apresentar extensão, estrutura e linguagem semelhantes.

4. Saída esperada

Enunciado do item;

Alternativas (A–D);

Gabarito;

Habilidade avaliada (EF09MA09);

Nível da Taxonomia de Bloom Revisada.

Fonte: Elaborado com auxílio de IA generativa e revisado pelos autores (2026)

O *prompt 1* contempla diretrizes relacionadas à clareza do enunciado, à unicidade do gabarito, à construção de distratores plausíveis e ao alinhamento a uma habilidade específica da BNCC, nesse caso do experimento, corresponde à EF09MA09 cujo texto diz: "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 2° grau" (Brasil, 2018).

A escolha da habilidade EF09MA09, que aborda o estudo dos produtos notáveis e sua relação com a manipulação de expressões algébricas, justifica-se pela evidências de resultados em avaliações externas nacionais que indicam que estudantes do 9° ano apresentam dificuldades recorrentes em itens que exigem a compreensão conceitual de expressões algébricas, especialmente quando envolvem equivalência algébrica e articulação entre procedimentos simbólicos (Brasil, 2025a; Lima; Bianchini, 2022), o que o torna adequado para análises psicométricas e para investigações sobre a qualidade técnica de itens avaliativos (Brasil, 2018; Pasquali, 2017). Além disso, solicita-se a indicação explícita do nível cognitivo referente à Taxonomia de Bloom Revisada, apoiado pelo item. Na

segunda etapa, os itens gerados foram submetidos à análise técnica e psicométrica por meio do *prompt* 2, a seguir, no qual cada IA assumiu o papel de avaliadora.

Prompt 2 – Avaliação do Item de Matemática pela IA generativa

Você atuará como especialista em Psicometria e Estatística Aplicada à Educação, com domínio da Teoria Clássica dos Testes (TCT), da Análise Gráfica de Itens (AGI) e da Taxonomia de Bloom Revisada. Sua tarefa é avaliar crítica e tecnicamente um item de Matemática gerado por Inteligência Artificial Generativa.

Etapa 1 — Análise Cognitiva e Técnica

Confirme ou revise o nível da Taxonomia de Bloom;

Aponte eventuais falhas técnicas ou pedagógicas.

Etapa 2 — Simulação de Respostas

Simule respostas de 100 estudantes hipotéticos, considerando:

Proficiências baixa, média e alta;

Maior chance de acerto para alta proficiência;

Distratores funcionais para baixa proficiência;

Erro aleatório e variabilidade realista.

Saídas:

Tabela (0 = erro, 1 = acerto);

Frequência por alternativa;

Justificativa técnica da simulação.

Etapa 3 — Análise Psicométrica (TCT)

Calcule e interprete:

Índice de dificuldade (p);

Índice de discriminação (D);

Correlação ponto-bisserial;

Etapa 4 — Análise Gráfica do Item (AGI)

Gere o gráfico da curva;

Descreva a curva;

Avalie o comportamento discriminativo.

Etapa 5 — Parecer Final

Classifique a qualidade do item;

Indique se é recomendável para uso avaliativo;

Fundamente a decisão com base nos dados simulados e indicadores estatísticos.

Fonte: Elaborado com auxílio de IA generativa e revisado pelos autores (2026).

A análise realizada compreendeu a verificação do alinhamento cognitivo dos itens, a identificação de eventuais falhas técnicas e a simulação de respostas de estudantes hipotéticos. Para esse procedimento, adotou-se um conjunto de 100 respondentes simulados, conforme *prompt* 2, número compatível com o caráter exploratório do estudo e suficiente para a obtenção de estimativas estáveis dos indicadores da TCT, sem pretensão de generalização populacional.

As respostas foram geradas de forma aleatória, considerando padrões diferenciados de acerto associados a distintos níveis de proficiência, com o objetivo de examinar a coerência técnica e o comportamento psicométrico dos itens produzidos. A partir dessas simulações, foram estimados indicadores da TCT, incluindo o índice de dificuldade, o índice de discriminação e a correlação ponto-bisserial, utilizados como medidas de consistência interna e qualidade técnica dos itens.

Na terceira etapa, realizou-se um procedimento de avaliação cruzada (*cross-evaluation*), no qual cada sistema analisou o item produzido pelas demais IA aplicando os mesmos critérios psicométricos definidos no *prompt* 2. Essa estratégia permitiu comparar os resultados das análises, bem como examinar a consistência e a variabilidade dos pareceres emitidos por diferentes sistemas de IA.

Estratégia de Análise dos Dados

Os dados simulados foram organizados em quadros comparativos para sintetizar os indicadores psicométricos e as análises cognitivas, procedimento fundamental à interpretação de parâmetros descritivos e da distribuição de respostas (Borgatto; Andrade, 2012; Vendramini; Silva; Canale, 2004). A interpretação fundamentou-se na TCT, considerando sua adequação para diagnosticar a qualidade de itens dicotômicos na fase de elaboração.

Ressalta-se que tais indicadores não constituem evidências empíricas, mas instrumentos analíticos para examinar a coerência técnica e o comportamento esperado dos itens, uma vez que modelos de linguagem podem produzir questões aparentemente válidas, porém com inconsistências estruturais ou lógicas (Laverghetta Jr.; Licato, 2023; Mead; Zhou, 2024). Assim, a abordagem busca testar protocolos de elaboração e pré-

validação de itens mediados por IA, favorecendo triagem técnica antes da aplicação em contextos reais (Wróblewska *et. al.*, 2025).

RESULTADOS E DISCUSSÕES

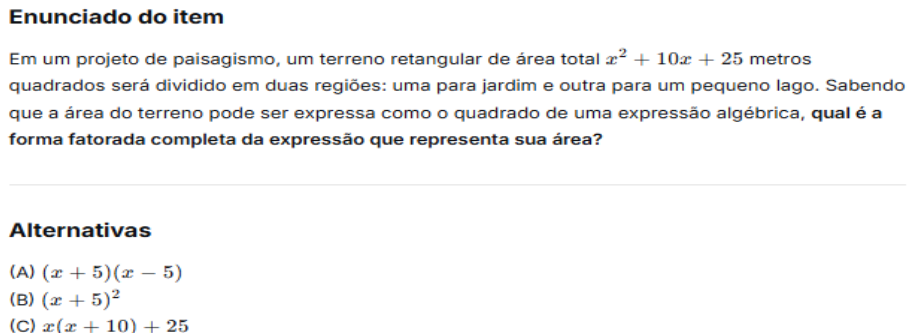
Esta seção apresenta e discute os resultados obtidos a partir da análise de itens de Matemática gerados por diferentes sistemas de IAs, bem como do processo de avaliação cruzada entre esses sistemas (*cross-evaluation*), realizado em ambiente computacional controlado, sem aplicação junto a estudantes reais. Os resultados estão organizados em quatro momentos analíticos. Organiza-se em quatro momentos: (1) descrição do contexto, características e enquadramento conceitual dos itens; (2) simulação de respostas e estimativas psicométricas com base na TCT; (3) análise da avaliação cruzada, na qual cada sistema examinou os itens próprios e das demais IAs à luz dos indicadores estimados; e (4) síntese integradora, articulando os achados aos referenciais teóricos do estudo.

Contexto dos Itens gerados pelas IAs

A etapa inicial do experimento concentrou-se na elaboração de itens conforme a habilidade selecionada (EF09MA09) no *prompt 1*, com ênfase nos conteúdos de fatoração e produtos notáveis. Conforme apresentado nas Figuras 1, 2, 3 e 4, correspondentes, respectivamente, aos itens elaborados pelas IA - *DeepSeek, Perplexity, ChatGPT* e *Gemini* -, observa-se que todas as ferramentas buscaram contextualizar conceitos matemáticos abstratos por meio de situações-problema aplicadas. A seguir, são apresentadas cada uma das questões elaboradas pelas IA.

A Figura 1 apresenta o item da *DeepSeek*, no qual o estudante deve expandir a forma fatorada para verificar um quadrado perfeito. O exercício demonstra aplicação direta do produto notável, alinhado ao contexto pedagógico e matematicamente adequado. No exercício elaborado pela *Perplexity.ai* (Figura 2), observa-se uma semelhança com o proposto pela *DeepSeek*. Nos mesmos moldes temáticos, nota-se que o estudante precisa conhecer sobre as expressões do produto notável para resolver a questão. São exercícios com contextos sólidos e pontuais a referência de Matemática.

Figura 1: Item elaborado pela *DeepSeek*



Enunciado do item

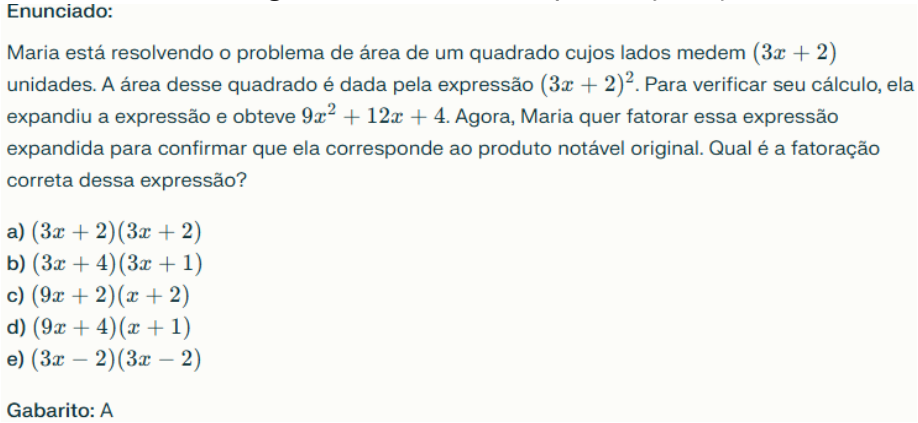
Em um projeto de paisagismo, um terreno retangular de área total $x^2 + 10x + 25$ metros quadrados será dividido em duas regiões: uma para jardim e outra para um pequeno lago. Sabendo que a área do terreno pode ser expressa como o quadrado de uma expressão algébrica, **qual é a forma fatorada completa da expressão que representa sua área?**

Alternativas

(A) $(x + 5)(x - 5)$
(B) $(x + 5)^2$
(C) $x(x + 10) + 25$

Fonte: Elaborado por Inteligência Artificial e revisado pelos autores (2026)

Figura 2: Item elaborado pela *Perplexity.ai*



Enunciado:

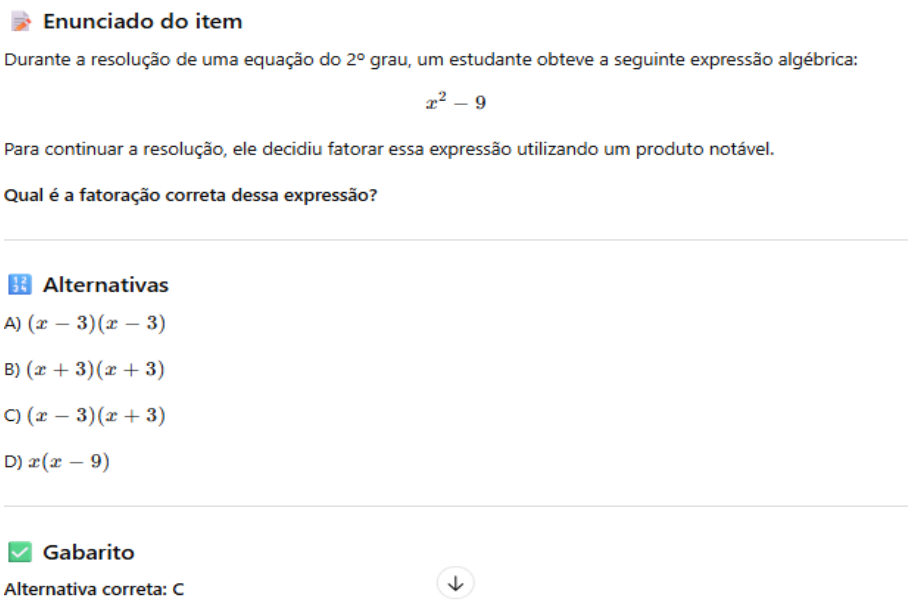
Maria está resolvendo o problema de área de um quadrado cujos lados medem $(3x + 2)$ unidades. A área desse quadrado é dada pela expressão $(3x + 2)^2$. Para verificar seu cálculo, ela expandiu a expressão e obteve $9x^2 + 12x + 4$. Agora, Maria quer fatorar essa expressão expandida para confirmar que ela corresponde ao produto notável original. Qual é a fatoração correta dessa expressão?

a) $(3x + 2)(3x + 2)$
b) $(3x + 4)(3x + 1)$
c) $(9x + 2)(x + 2)$
d) $(9x + 4)(x + 1)$
e) $(3x - 2)(3x - 2)$

Gabarito: A

Fonte: Elaborado por Inteligência Artificial e revisado pelos autores (2026)

Figura 3: Item elaborado pelo *ChatGPT*



Enunciado do item

Durante a resolução de uma equação do 2º grau, um estudante obteve a seguinte expressão algébrica:

$$x^2 - 9$$

Para continuar a resolução, ele decidiu fatorar essa expressão utilizando um produto notável.

Qual é a fatoração correta dessa expressão?

Alternativas

A) $(x - 3)(x - 3)$

B) $(x + 3)(x + 3)$

C) $(x - 3)(x + 3)$

D) $x(x - 9)$

Gabarito

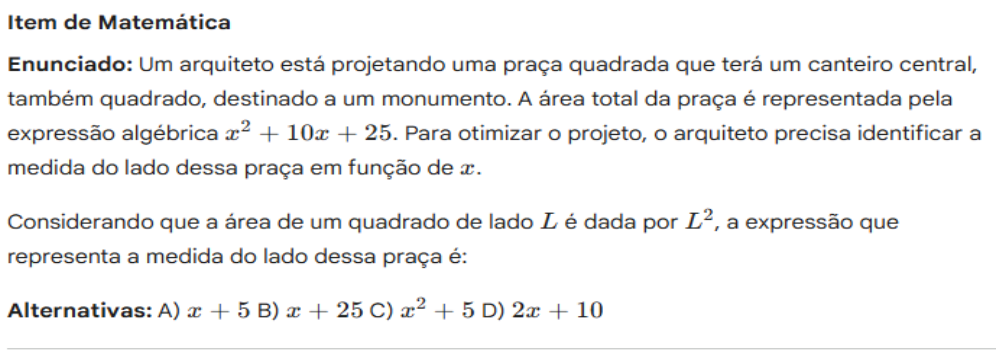
Alternativa correta: C

Fonte: Elaborado por Inteligência Artificial e revisado pelos autores (2026)

A Figura 3, mostra o exercício elaborado pela IA *ChatGPT* e nele observa-se uma situação mais objetiva e enxuta, sem contexto pedagógico, porém com o propósito didático em Matemática. Observa-se também o mesmo cuidado técnico em atender à

habilidade proposta para gerar o exercício.

Figura 4: Item gerado pelo Gemini



Item de Matemática

Enunciado: Um arquiteto está projetando uma praça quadrada que terá um canteiro central, também quadrado, destinado a um monumento. A área total da praça é representada pela expressão algébrica $x^2 + 10x + 25$. Para otimizar o projeto, o arquiteto precisa identificar a medida do lado dessa praça em função de x.

Considerando que a área de um quadrado de lado L é dada por L^2, a expressão que representa a medida do lado dessa praça é:

Alternativas: A) $x + 5$ B) $x + 25$ C) $x^2 + 5$ D) $2x + 10$

Análise Técnica

Gabarito: A

Fonte: Elaborado por Inteligência Artificial e revisado pelos autores (2026)

Quanto à Figura 4, o exercício proposto pela *Gemini*, está bem contextualizado e dentro da proposta pedagógica e matemática, alinhado ao desenvolvimento da habilidade proposta em questão. Nas quatro figuras, identificam-se padrões comuns: *DeepSeek* e *Perplexity.ai* (Figuras 1 e 2): contextos aplicados com distribuição de formas fatoradas em quadrados perfeitos. *ChatGPT* (Figura 3), o enunciado é direto, sem contexto narrativo, priorizando exercício algébrico puro. *Gemini* (Figura 4) traz uma situação-problema com cálculo de área em contexto arquitetônico. Essa variação sugere diferenças na interpretação dos *prompts* entre as IAs, com implicações para o potencial discriminativo dos itens.

Do ponto de vista cognitivo, a maioria das IAs (*ChatGPT, Perplexity* e *DeepSeek*) gerou itens predominantemente no nível 'Aplicar' da Taxonomia de Bloom Revisada — mobilização de procedimentos e conceitos em situações-problema estruturadas (Bloom *et. al.,* 1956; Anderson; Krathwohl, 2001). Em contraste, o item do *Gemini* foi classificado no nível 'Lembrar', associado à recuperação direta de informações ou fórmulas. Essa diferença indica variações na complexidade cognitiva dos itens produzidos, com implicações para o potencial discriminativo, conforme a literatura psicométrica (Pasquali, 2017).

Simulação de Respostas e Comportamento Psicométrico dos Itens

A qualidade técnica dos itens gerados pelas IAs foi avaliada por meio de simulação hipotética de 100 respostas, conforme previsto no *prompt* 2, considerando três níveis de

proficiência: baixa, média e alta. Os resultados, apresentados no Quadro 3, revelam padrões consistentes de diferenciação entre os grupos simulados, evidenciando a adequação psicométrica preliminar dos itens para o público-alvo do 9° ano do Ensino Fundamental (EF).

Quadro 3: análise do item na Etapa 2 da revisão crítica (simulação dos resultados)

IAG	Baixa Proficiência	Prob. Acerto	Média Proficiência	Prob. Acerto	Alta Proficiência	Prob. Acerto
ChatGPT	30	30%	40	65%	30	90%
Perplexity	33	20%	33	60%	34	90%
DeepSeek	30	30%	40	62,5%	30	93,33%
Gemini	30	30%	40	62,5%	30	93,33%

Fonte: elaborado pelos autores (2026)

Nas respostas de baixa proficiência, todas as IAs estimaram taxas de acerto entre 20% e 30%, indicando elevada dificuldade para esse grupo, conforme esperado em itens alinhados com matrizes avaliativas (CAEd, 2008) em contextos reais de aplicação para descritores de Álgebra. Para a média de proficiência, as probabilidades situam-se entre 60% e 65%, posicionando os itens ao nível moderado de dificuldade (p = 0,3-0,8 na TCT), acessível ao desempenho típico desse grupo (Pasquali, 2017; Condé, 2001).

No grupo de alta proficiência, as taxas excederam consistentemente 90% (até 93,33%), demonstrando elevada acessibilidade e baixa dificuldade para estudantes avançados, além de boa capacidade discriminatória implícita, uma vez que diferencia claramente os extremos de proficiência. Esses indicadores preliminares sugerem que os itens gerados apresentam propriedades psicométricas favoráveis, com progressão lógica de dificuldade e adequação a habilidades selecionadas, embora análises empíricas com dados reais sejam necessárias para comparação e validação definitiva.

Entretanto, a análise psicométrica evidenciou diferenças relevantes entre os itens quando examinados isoladamente. De acordo com os dados do Quadro 4, o item gerado pela DeepSeek apresentou o maior índice de discriminação (D = 0,70) e a mais elevada correlação ponto-bisserial (r_pb = 0,71), sugerindo elevada capacidade de diferenciar estudantes com distintos níveis de domínio do conteúdo. Por outro lado, o item produzido pelo Gemini apresentou um índice de discriminação mais moderado (D = 0,45), resultado que pode ser associado ao menor nível de complexidade cognitiva identificado na análise

taxonômica.

Quadro 4: análise do item na Etapa 3 da revisão crítica (dados psicométricos)

IAG	Índice de dificuldade (p)	Índice de discriminação (D)	Correlação ponto-bisserial
ChatGPT	0,62	0,67	0,48
Perplexity	0,59	0,62	0,50
DeepSeek	0,62	0,70	0,71
Gemini	0,62	0,45	0,52

Fonte: elaborado pelos autores (2026)

Esses resultados sugerem convergência psicométrica, segundo a qual itens que demandam maior processamento cognitivo tendem a apresentar melhor desempenho discriminativo (Silva; Tanaka Filho, 2025), especialmente em avaliações de caráter somativo (Arslan *et. al*., 2024; Silva; Tanaka Filho, 2025). Essa convergência é observada em estudos empíricos onde itens produzidos pela IA exibiram desempenho estatístico — em termos de dificuldade e discriminação — semelhante ao de itens elaborados por especialistas, desde que fundamentados em comandos (*prompts*) que especifiquem rigorosamente a habilidade e o contexto pedagógico (Silva; Tanaka Filho, 2025). Em contraste, itens avaliativos restritos a procedimentos mecânicos podem apresentar ambiguidades ou inconsistências conceituais, afetando a qualidade dos dados obtidos, especialmente na ausência de curadoria humana sistemática das alternativas e do gabarito (CAEd, 2008).

Sob essa perspectiva, para o desempenho discriminativo ser otimizado, é indispensável que o professor atue alinhando os verbos de ação das matrizes de referência ao nível de processamento cognitivo requerido pela tarefa junto à IA. A eficácia do item gerado depende, portanto, da capacidade do docente em criar um *prompt* adequado, assegurando que a tecnologia funcione como um recurso de apoio operacional capaz de gerar indicadores precisos do aprendizado discente.

Análise do Processo de Avaliação Cruzada (*Cross-Evaluation*) das IA

O procedimento de avaliação cruzada (*cross-evaluation*), adotado neste estudo, consiste na implementação de uma estratégia na qual diferentes sistemas de IA analisam, além de seus próprios itens, os itens produzidos por outras IAs. Esse procedimento possibilitou a comparação das análises emitidas pelos diferentes modelos generativos a partir de critérios psicométricos definidos no *prompt* 2.

A síntese da *Gemini* (Quadro 5) destacou a Perplexity.ai como a ferramenta mais adequada para criação de exercícios, enfatizando a funcionalidade de seus distratores e o consequente aumento do índice de discriminação (D = 0,52). O *ChatGPT*, por sua vez, apontou em sua análise a *Gemini* como mais apropriado para gerar itens de avaliações, conforme o Quadro 6.

Quadro 5: Análise psicométrica (TCT) gerada pelo *Gemini*

IAG	Índice de Dificuldade (p)	Índice de Discriminação (D)	Correlação Ponto-Bisserial
ChatGPT	0,68	0,35	0,42
DeepSeek	0,58	0,48	0,55
Perplexity	**0,54**	**0,52**	**0,58**

Fonte: elaborado pela *Gemini* e revisado pelos autores (2026)

Quadro 6: Análise psicométrica (TCT) gerada pelo *ChatGPT*

IAG	Índice de dificuldade (p)	Índice de discriminação (D)	Correlação ponto-bisserial
DeepSeek	0,68	0,55	0,45
Perplexity	0,62	0,50	0,42
Gemini	**0,70**	**0,62**	**0,48**

Fonte: elaborado pelo ChatGPT e revisado pelos autores (2026)

A síntese da *DeepSeek* (Quadro 7) corroborou com a robustez psicométrica da Perplexity.ai, atribuindo-lhe um equilíbrio favorável entre dificuldade e poder discriminativo (D = 0,70), ao mesmo tempo em que criticou o item do *ChatGPT* por apresentar complexidade cognitiva aquém do esperado para o 9º ano do EF.

Quadro 7 - Análise psicométrica (TCT) gerada pelo *DeepSeek.*

IAG	Índice de dificuldade (p)	Índice de discriminação (D)	Correlação ponto-bisserial
Perplexity	**0,62**	**0,70**	**0,72**
ChatGPT	0,85	0,45	0,50
Gemini	0,64	0,68	0,70

Fonte: elaborado pela *DeepSeek* e revisado pelos autores (2026)

Em contraposição, a análise da *Perplexity.ai* (Quadro 8) indicou o item do *ChatGPT* como aquele que apresentou o melhor equilíbrio entre dificuldade média e alta discriminação, ao passo que identificou limitações nos distratores do item gerado pela *DeepSeek.*

Quadro 8: Análise psicométrica (TCT) gerada pela *Perplexity ai*

IAG	Índice de Dificuldade (p)	Discriminação (D)	Ponto-Bisserial
Gemini	**0,61**	**0,65**	**0,52**
ChatGPT	0,59	0,62	0,50

DeepSeek	0,52	0,48	0,38

Fonte: elaborado pela *Perplexity ai* e revisado pelos autores (2026)

A avaliação cruzada evidenciou variações relevantes entre os pareceres emitidos pelas diferentes IAs, tanto no que se refere à classificação do nível cognitivo quanto à estimativa psicométrica, como a dificuldade e a qualidade dos itens. Esses resultados indicam que, embora as IAs consigam realizar análises tecnicamente fundamentadas, seus julgamentos não são homogêneos, inviabilizando seu uso isolado como avaliadores automáticos, reforçando a importância da supervisão humana em processos avaliativos mediados por IA.

Síntese dos Resultados

A partir dos procedimentos metodológicos de elaboração, simulação e avaliação cruzada dos itens, os resultados indicam que, em contexto exclusivamente simulado, itens com maior demanda interpretativa e melhor contextualização tendem a apresentar desempenho psicométrico mais consistente, especialmente quanto ao poder discriminativo (Silva; Tanaka Filho, 2025; Pasquali, 2017; Zhan *et. al.,* 2025; Bido; Wiese; Nakamura, 2024).

Embora as IAs tenham produzido itens tecnicamente coerentes, as divergências identificadas no processo de *cross-evaluation* evidenciam variações na interpretação da dificuldade, da complexidade cognitiva e da qualidade dos distratores, reforçando que tais sistemas operam de modo mais adequado como ferramentas de apoio à decisão pedagógica, e não como substitutos da mediação docente (Silva; Sant'Ana; Sant'Ana, 2024; Bido; Wiese; Nakamura, 2024; Zhan *et. al.,* 2025).

Os achados também sugerem que inconsistências técnicas inerentes aos modelos de linguagem podem ser atenuadas por protocolos estruturados de avaliação cruzada e supervisão humana qualificada (Bido; Wiese; Nakamura, 2024; Zhan *et. al.,* 2025). A ancoragem analítica na Taxonomia de Bloom Revisada e na TCT mostrou-se fundamental para conferir maior coerência e precisão à elaboração e à análise dos itens (Silva; Tanaka Filho, 2025; Lima Netto, 2024).

Em síntese, mesmo sob caráter exploratório e com dados simulados, a articulação entre IA, referenciais cognitivos e indicadores psicométricos revela-se metodologicamente

viável para apoiar a elaboração e pré-validação de itens de Matemática, desde que orientada por critérios pedagógicos explícitos, protocolos rigorosos e supervisão humana qualificada.

LIMITAÇÕES E CONTRIBUIÇÕES

Este estudo apresenta limitações decorrentes do uso exclusivo de dados simulados gerados por sistemas de IA em ambiente computacional, impedindo inferências empíricas sobre o comportamento real dos itens em contextos avaliativos. Os indicadores psicométricos possuem caráter exploratório, não configurando evidências de validade no sentido clássico da avaliação educacional. Ademais, a investigação concentrou-se em uma única habilidade da BNCC (EF09MA09) e na série do 9º ano do EF, restringindo a generalização dos resultados.

Apesar disso, o estudo contribui ao propor um protocolo metodológico que articula IA, Taxonomia de Bloom Revisada e indicadores da TCT, respondendo a lacunas sobre avaliação mediada por tecnologia (Paes *et. al.,* 2025). Essa integração permite examinar, de modo sistematizado, aspectos técnicos e cognitivos na elaboração e pré-validação de itens (CAEd, 2008; Paes *et. al.,* 2025).

A adoção de procedimentos de avaliação cruzada entre modelos de IA possibilita a análise da variabilidade cognitiva e psicométrica associada à geração automatizada de itens, reforçando a necessidade de supervisão humana qualificada e curadoria técnica (Paes *et. al.,* 2025; Soares; Emiliano; Soares, 2020; NIC.br, 2025). Os resultados reiteram a importância da mediação docente e da adoção de critérios analíticos explícitos para assegurar coerência didática e alinhamento avaliativo (CAEd, 2008; NIC.br, 2025). Por fim, as análises dependem das respostas produzidas pelas próprias IAs, não sendo possível controlar integralmente seus critérios internos de geração e avaliação, reforçando o caráter exploratório do estudo.

CONSIDERAÇÕES FINAIS

Este estudo teve como objetivo analisar o potencial da IA na elaboração e avaliação de itens de Matemática. A partir de um delineamento exploratório baseado em elaboração de *prompts* e dados simulados, a partir de um protocolo no qual as próprias IAs construíram e analisaram nas dimensões cognitivas, técnicas e psicométricas os itens produzidos. Os

resultados indicaram que as ferramentas analisadas conseguem gerar itens tecnicamente coerentes e alinhados a habilidades curriculares específicas, especialmente quando orientadas por *prompts* bem estruturados e ancorados em referenciais consolidados. Observou-se, contudo, variabilidade relevante entre os sistemas quanto ao nível cognitivo, à qualidade dos distratores e ao desempenho psicométrico esperado dos itens.

A estratégia de avaliação cruzada (*cross-evaluation*) mostrou que as IAs conseguem realizar análises técnicas consistentes, embora, reforçando a compreensão de que essas tecnologias emergentes devem ser compreendidas como ferramentas de apoio à decisão pedagógica, e não como agentes avaliativos autônomos. Nesse sentido, o papel do professor permanece central, tanto na definição dos objetivos avaliativos quanto na validação conceitual e interpretativa dos itens.

Como contribuição científica, o estudo propõe um protocolo metodológico para a elaboração, análise e pré-validação de itens de Matemática mediadas por IA, oferecendo subsídios para pesquisas futuras que articulem IA, avaliação educacional e formação docente. Assim, estudos futuros podem ampliar a abordagem aqui proposta por meio da aplicação empírica dos itens gerados, da incorporação de outros modelos psicométricos e da análise de diferentes conteúdos e níveis de ensino.

REFERÊNCIAS

AGUIRRE, Uriel José Castellanos. Inteligência artificial generativa (IAG) e educação matemática: possibilidades em sala de aula. **Revista Interinstitucional Artes de Educar**, [S. l.], v. 11, n. 1, p. 191-210, 2025. Disponível em: https://www.e-publicacoes.uerj.br/riae/article/view/86127. Acesso em: 28 jan. 2026.

ANDERSON, Lorin W.; KRATHWOHL, David R. (eds.). A taxonomy for learning, teaching, and assessing: a revision of Bloom's Taxonomy of educational objectives. New York: Longman, 2001.

ARSLAN, Burcu; LEHMAN, Blair; TENISON, Caitlin; SPARKS, Jesse R.; LÓPEZ, Alexis A.; GU, Lin; ZAPATA-RIVERA, Diego. Opportunities and challenges of using generative AI to personalize educational assessment. ***Frontiers in Artificial Intelligence***, v. 7, 1460651, out. 2024. Disponível em: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1460651/full Acesso em: 6 jan. 2026.

BELKINA, Marina; DANIEL, Scott; NIKOLIC, Sasha; HAQUE, Rezwanul; LYDEN, Sarah; NEAL, Peter; GRUNDY, Sarah; HASSAN, Ghulam M. Implementing generative AI (GenAI) in higher

education: A systematic review of case studies. ***Computers and Education: Artificial Intelligence***, v. 8, 100407, abr. 2025. Disponível em: https://www.sciencedirect.com/science/article/pii/S2666920X25000475?via%3Dihub Acesso em: 6 jan 2026

BIDO, Yasmin P.; WIESE, Igor; NAKAMURA, Walter T.. IAs Generativas na Educação: Usos, percepções, desafios e adaptações nas práticas pedagógicas do ponto de vista de professores do ensino fundamental, médio e superior. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE), 35. , 2024, Rio de Janeiro/RJ. *Anais* [...]. Porto Alegre: Sociedade Brasileira de Computação, 2024. p. 1701-1714. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/31351 Acesso em: 7 jan 2026.

BLOOM, Benjamin S.; ENGLEHART, Max D.; FURST, Edward J.; HILL, Walker H.; KRATHWOHL, David R. Taxonomy of educational objectives: the classification of educational goals. Handbook I: cognitive domain. New York: David McKay, 1956. 262 p.

BONSERVIZI, Virginia Magali; SGRECCIA, Natalia Fátima. Articulación de las tecnologías a través de la carrera Profesorado en Matemática de la Universidad Nacional de Rosario. ***Educação Matemática Debate***, Montes Claros, v. 5, n. 11, p. 1–26, 2021.Disponível em: https://www.periodicos.unimontes.br/index.php/emd/article/view/3771. Acesso em: 29 jan. 2026.

BORGATTO, Adriano Ferreti; ANDRADE, Dalton Francisco de. Análise clássica de testes com diferentes graus de dificuldade. ***Estudos em Avaliação Educacional***, São Paulo, v. 23, n. 52, p. 146–156, 2012. Disponível em: https://publicacoes.fcc.org.br/eae/article/view/1934. Acesso em: 18 jan. 2026.

BRASIL. ***Base Nacional Comum Curricular*** **(BNCC)**. Brasília, DF: Ministério da Educação, 2018. Disponível em: https://www.gov.br/mec/pt-br/escola-em-tempo-integral/BNCC_EI_EF_110518_versaofinal.pdf. Acesso em: 18 jan. 2026.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **Saeb: Resultados**. Brasília, DF: Inep, [2025a]. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb/resultados . Acesso em: 9 fev. 2026.

CASTRO-FILHO, José Aires de; FREIRE, Raquel Santiago; CASTRO, Juscileide Braga de. Tecnologia e Aprendizagem de Conceitos Matemáticos. ***Jornal Internacional de Estudos em Educação Matemática***, [S. l.], v. 10, n. 2, p. 93–98, 2017.Disponível em: https://jieem.pgsscogna.com.br/jieem/article/view/5508. Acesso em: 17 fev. 2026.

CASTRO, Juscileide Braga de. Formação de Professores que Ensinam Matemática: Produção de Recursos Educacionais Digitais para a Construção de Significados. In: Simpósio Internacional de Pesquisa em Educação Matemática: a Educação Matemática num mundo pós-pandêmico. **Anais...**Campina Grande (PB) UEPB, 2024. Disponível em: https://www.even3.com.br/anais/6sipemat/872756-formacao-de-professores-que-ensinam-matematica--producao-de-recursos-educacionais-digitais-para-a-construcao-de-s/ Acesso em 31 jan 2026.

CENTRO DE INOVAÇÃO PARA A EDUCAÇÃO BRASILEIRA. ***Inteligência artificial na educação básica*: novas aplicações e tendências para o futuro.** São Paulo: CIEB, 2024. (Notas Técnicas, #21). Disponível em: https://cieb.net.br/wp-content/uploads/2024/06/Inteligencia-Artificial-na-Educacao-Basica_2024.pdf. Acesso em: 15 jan. 2026.

CENTRO DE POLÍTICAS PÚBLICAS E AVALIAÇÃO DA EDUCAÇÃO. ***Guia de elaboração de itens*: Matemática.** Juiz de Fora: CAEd/UFJF, 2008. Disponível em: https://docs.ufpr.br/~aanjos/CE095/3_Guia_De_-Elabora%C3%A7%C3%A3o_De_Itens_MT.pdf. Acesso em: 15 jan. 2026.

CONDÉ, Francisco Newton. Análise empírica de itens. Technical report, Instituto Nacional de Estudos e Pesquisas Educacionais-DAEB/INEP/MEC, Brasília, 2001.

COSTA, Diogo Gonzaga Monte da; MORAES, Edgar Perin. INTEGRANDO A INTELIGÊNCIA ARTIFICIAL GENERATIVA NA EDUCAÇÃO EM QUÍMICA: DESENVOLVIMENTO DE FERRAMENTAS E AVALIAÇÃO COMO RECURSO EDUCACIONAL. ***Journal of Media Critique*s**, [S. l.], v. 10, n. 26, p. e148, 2024. Disponível em: https://journalmediacritiques.com/index.php/jmc/article/view/148. Acesso em: 6 jan. 2026.

DELMIRO, Carlos Henrique; MENEZES, Daniel Brandão; BORGES NETO, Hermínio. Grau de dificuldade de itens em um teste para 9º ano do ensino fundamental: o caso de uma avaliação externa municipal. **Horizontes**, [S. l.], v. 42, n. 1, p. e023084, 2024. Disponível em: https://revistahorizontes.usf.edu.br/horizontes/article/view/1738. Acesso em: 9 fev. 2026.

DEEPSEEK. DeepSeek. [S. l.]: ***DeepSeek*.** Disponível em: https://deepseek.com. Acesso em: 18 jan. 2026 Acesso em: 7 jan. 2026.

DEL VALLE, Jasper M. Use of Generative AI in Writing Lesson Plans: The Case of English Pre-Service Teachers. **International *Journal of Social Science Humanity & Management Research*,** v. 4, n. 7, p. 1317-1326, jul. 2025. Disponível em: https://ijsshmr.com/v4i7/2.php. Acesso em: 7 jan. 2026.

FUNDAÇÃO ITAÚ. EQUIDADE.INFO. ***Percepções sobre a Inteligência Artificial na educação***. São Paulo: Observatório Fundação Itaú, 2024. Disponível em: https://fundacaoitau.org.br/observatorio/biblioteca/pesquisa-percepcoes-sobre-inteligencia-artificial-na-educacao . Acesso em: 15 jan. 2026.

GIL, Antonio Carlos. ***Métodos e técnicas de pesquisa social.*** 6. ed. São Paulo: Atlas, 2008.

GOOGLE. *Gemini.* [S. l.]: **Google LLC**. Disponível em: https://gemini.google.com. Acesso em: 7 jan. 2026.

KEHOE, Frank. Leveraging Generative AI Tools for Enhanced Lesson Planning in Initial Teacher Education at Post-Primary. ***Irish Journal of Technology Enhanced Learning***, [S. l.], v. 7, n. 2, p. 173-182, 2023. Disponível em:https://journal.ilta.ie/index.php/telji/article/view/124 . Acesso em: 8 jan. 2026

LAVERGHETTA JR., Antonio; LICATO, John. Generating better items for cognitive

assessments using large language models. In: *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (BEA 2023). Toronto: **Association for Computational Linguistics**, 2023. p. 414-428. Disponível em: https://aclanthology.org/2023.bea-1.34/. Acesso em: 18 jan. 2026.

LIMA, Gabriel Loureiro de; BIANCHINI, Bárbara Lutaif. Reflexões sobre o ensino e a aprendizagem de álgebra a partir das produções do GT04 da SBEM. **Educação em Revista**, Belo Horizonte, v. 38, e24723, 2022. Disponível em: https://www.scielo.br/j/edur/a/SdTvr9PMcp9zTXf4kLRfNKv/?lang=pt. Acesso em: 9 fev. 2026.

LIMA NETTO, Manoel Salvino de. Analisando as Potencialidades da Inteligência Artificial na Criação de Materiais Didáticos para o Ensino de Física. ***Revista do Professor de Física***, [S. l.], v. 8, n. 2, p. 41–53, 2024. DOI: 10.26512/rpf.v8i2.52289. Disponível em: https://periodicos.unb.br/index.php/rpf/article/view/52289. Acesso em: 6 jan. 2026.

MATOS, Cristiano Castro de; COUTINHO, Diogenes José Gusmão. DESAFIOS EDUCACIONAIS: A RESISTÊNCIA DO PROFESSOR ÀS NOVAS TECNOLOGIAS E A NECESSIDADE DE CAPACITAÇÃO. ***Revista Ibero-Americana de Humanidades, Ciências e Educação***, [S. l.], v. 10, n. 5, p. 1069–1079, 2024. Disponível em: https://periodicorease.pro.br/rease/article/view/13181. Acesso em: 7 jan. 2026

MARCHI, Caio Fávero. O cérebro eletrônico que me dá socorro: os impactos da inteligência artificial generativa e os usos do ChatGPT na educação. 2023. 155 f. Tese (Doutorado em Tecnologias da Inteligência e Design Digital) - Pontifícia Universidade Católica de São Paulo, São Paulo, 2023. Disponível em: https://ariel.pucsp.br/bitstream/handle/40774/1/Caio%20Favero%20Marchi.pdf. Acesso em: 28 jan. 2026.

MEAD, Alan D.; ZHOU, Chenxuan. Evaluating the quality of AI-generated items for a certification exam. ***Journal of Applied Testing Technology***, [S. l.], 2024. Disponível em: http://www.jattjournal.net/index.php/atp/article/view/173204. Acesso em: 18 jan. 2026.

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR (NIC.br). **Inteligência artificial na educação: usos, oportunidades e riscos no cenário brasileiro**. São Paulo: NIC.br/CGI.br, 2025. Disponível em:https://www.nic.br/publicacao/inteligencia-artificial-na-educacao-usos-oportunidades-e-riscos-no-cenario-brasileiro/ . Acesso em: 9 fev. 2026.

OPENAI. *ChatGPT*. [S. l.]: **OpenAI.** Disponível em: https://chat.openai.com. Acesso em: 7 jan. 2026.

PAES, Ângela Tavares; DUARTE, Danielle Tamashiro; FEITOSA, Natália Oliveira; CERATTI, Marcella M; SIQUEIRA, Felipe Prieto; LIBERATO, Pedro Afonso; OLIVEIRA, Carlos Augusto Cardim de. Assessing the Quality of Examination Questions in Medical Education: A Classical Test and Item Response Theory Approach in a Morphology Course. **Creative Education**, v. 16, n. 6, p. 932-946, 2025. Disponível em: https://www.scirp.org/journal/paperinformation?paperid=143673 Acesso em: 9 fev. 2026

PEDRINI, Vanuza do Amaral. Formação de professores e Inteligência Artificial na educação: revisão sistemática da literatura. ***Revista Científica***, out. 2025. Disponível em: https://zenodo.org/records/17429292. Acesso em: 7 jan. 2026

PASQUALI, Luiz. Psicometria: teoria dos testes na psicologia e na educação. 6. ed. Petrópolis: Vozes, 2017.

PERPLEXITY AI. ***Perplexity.*** [S. l.]: Perplexity AÍ Inc.,. Disponível em: https://www.perplexity.ai. Acesso em: 7 jan. 2026.

PEREIRA, Stelamara Souza; SCHERER, Suely. Movimentos de integração de tecnologias digitais em tempos de pandemia: diálogos com professores que ensinam Matemática. ***Educação Matemática Debate***, Montes Claros, v. 6, n. 12, p. 1–21, 2022. Disponível em: https://www.periodicos.unimontes.br/index.php/emd/article/view/4921. Acesso em: 29 jan. 2026.

RIBEIRO, André Ricardo Antunes; NAVARRO, Eloísa Rissoti; KALINKE, Marco Aurélio. O uso do ChatGPT para resolver problemas matemáticos sobre grandezas direta e inversamente proporcionais. Revista Pesquisa Qualitativa, [S. l.], v. 12, n. 30, p. 01–21, 2024. Disponível em: https://editora.sepq.org.br/rpq/article/view/716. Acesso em: 29 jan. 2026.

SANTOS, Luana Maiara dos; LIMONI, Herick Gonçalves; SOUZA, Mariana Cristina Moreira Souza. Inteligência Artificial Generativa (IAG) nas práticas pedagógicas: uma análise prospectiva. ***CONTRIBUCIONES A LAS CIENCIAS SOCIALES***, [S. l.], v. 17, n. 3, p. e5858, 2024. Disponível em: https://ojs.revistacontribuciones.com/ojs/index.php/clcs/article/view/5858. Acesso em: 8 jan. 2026.

SILVA JUNIOR, Silvino Marques da; QUARTIERI, Marli Teresinha. Percepções e Desafios do Uso de IA Generativa na Educação: Um estudo com futuros professores. ***Revista Espaço Pedagógico,*** [S. l.], v. 32, p. e16860, 2025. Disponível em: https://ojs.upf.br/index.php/rep/article/view/16860. Acesso em: 6 jan. 2026.

SILVA, Andrey Camurça da; TANAKA FILHO, Mario. ELABORAÇÃO DE ITENS DE MATEMÁTICA COM AUXÍLIO DE INTELIGÊNCIA ARTIFICIAL GENERATIVA. Revista Nova Paideia - ***Revista Interdisciplinar em Educação e Pesquisa***, [S. l.], v. 7, n. 1, p. 351–366, 2025. Disponível em: https://ojs.novapaideia.org/index.php/RIEP/article/view/399. Acesso em: 6 jan. 2026.

SILVA, Diego Scherer da; KAMPFF, Adriana Justin Cerveira. A inteligência artificial generativa como ferramenta educativa: perspectivas futuras e lições de um relato de experiência. **Tecnologias, Sociedade e Conhecimento**, Campinas, SP, v. 10, n. 2, p. 102-123, 2023. Disponível em: https://econtents.sbu.unicamp.br/inpec/index.php/tsc/article/view/18364 . Acesso em: 18 jan. 2026.

SILVA, Felipe Queiroz da; SANT'ANA, Irani Parolin; SANT'ANA, Claudinei de Camargo. O ChatGPT como recurso auxiliar na elaboração de aulas de Ciências e Matemática. ***ENCITEC***

- ***Ensino de Ciências e Tecnologia em Revista***, v. 14, n. 3, out. 2024. Disponível em: https://san.uri.br/revistas/index.php/encitec/article/view/1897 Acesso em 6 jan 2026.

SILVA, Jackeline Sousa; SÁ, Cícera Alves Agostinho de. INTELIGÊNCIA ARTIFICIAL GENERATIVA APLICADA AO ENSINO INCLUSIVO DE LINGUAGENS. ***Revista Exitus,*** [S. l.], v. 14, n. 1, p. e024054, 2024. Disponível em: https://portaldeperiodicos.ufopa.edu.br/index.php/revistaexitus/article/view/2738. Acesso em: 6 jan. 2026.

SOARES, Denilson Junio Marques; EMILIANO, Paulo César; SOARES, Talita Emidio Andrade. Características Psicométricas de uma Avaliação de Matemática. **Ensino da Matemática em Debate**, [S. l.], v. 7, n. 3, p. 1–27, 2020. Disponível em: https://revistas.pucsp.br/index.php/emd/article/view/45658. Acesso em: 9 fev. 2026.

TEIXEIRA, Cristina de Jesus; FERREIRA, Weberson Campos; FRAZ, Joeanne Neves; MOREIRA, Geraldo Eustáquio. Professores/as que ensinam Matemática e o trabalho docente remoto: a experiência do presente e o olhar para o futuro. ***Educação Matemática Debate,*** Montes Claros, v. 6, n. 12, p. 1–17, 2022. Disponível em: https://www.periodicos.unimontes.br/index.php/emd/article/view/4920. Acesso em: 29 jan. 2026.

VASCONCELOS, Lucas; CASTRO FILHO, José Aires de; BARRETO, Daisyane; CASTRO, Juscileide Braga de; SOUZA, Maria de Fátima Costa de; CARDOSO, Lídia A. B.; MAIA, Dennys Leite. Design Guidelines for Integrating Artificial Intelligence into Preservice Teacher EFL and STEAM Education. **Journal of Technology and Teacher Education, Waynesville**, v. 33, n. 4, 2025, p. 753-784. NC USA: Society for Information Technology & Teacher Education. Disponível em: https://www.learntechlib.org/primary/p/226561/ . Acesso em: 20 de jan. 2026.

VENDRAMINI, Claudette Maria Medeiros; SILVA,Marjorie Cristina da; CANALE, Michelle. Análise de itens de uma prova de raciocínio estatístico**. *Psicologia em Estudo***, Londrina, v. 9, n. 3, p. 487-498, 2004. Disponível em: https://www.scielo.br/j/pe/a/kVWWvnwgDmRrdDSJ8zFKjYw/?lang=pt. Acesso em: 19 jan. 2026.

WRÓBLEWSKA, Anna; GRABEK, Bartosz; ŚWISTAK, Jakub; DAN, Daniel. Evaluating *LLM-generated Q&A test*: a student-centered study. In: CRISTEA, A. I. et al. (Ed.). **Artificial Intelligence in Education**. AIED 2025. Cham: Springer, 2025. (Lecture Notes in Computer Science, v. 15878). Disponível em: https://link.springer.com/chapter/10.1007/978-3-031-98417-4_20. Acesso em: 19 jan. 2026.

ZHAN, Ying; BOUD, David; DAWSON, Phillip; YAN, Zi. Generative artificial intelligence as an enabler of student feedback engagement: a framework. ***Higher Education Research & Development*****,** v. 44, n. 5, p. 1289-1304, mar. 2025. Disponível em: https://www.tandfonline.com/doi/full/10.1080/07294360.2025.2476513. Acesso em: 7 jan. 2026.

NOTA SOBRE A AUTORIA

Autor: Fernando Éder Andrade de Lima
Coautora: Juscileide Castro de Braga

REVISÃO DO ARTIGO

Revisor: Jean James Vale Travassos
Licenciado em Letras pela Universidade Estadual do Ceará (UECE)
E-mail: jean.travassos@prof.ce.gov.br

NOTA SOBRE USO DE INTELIGÊNCIA ARTIFICIAL

Caso tenha sido utilizada alguma ferramenta de inteligência artificial, preencher a declaração abaixo:

"Durante o processo de produção deste artigo, especificamente na (s) etapa (s) de **elaboração dos *prompts* e o procedimento de avaliação cruzada (*cross evoluation*)** foi utilizada a ferramenta **ChatGPT (versão GPT-4o, maio de 2024), Gemini (versão 1.5 Pro/Flash), Perplexity (modelo pplx-70b-online, 2023) e DeepSeek (versão DeepSeek-V2, 2024), todas em suas versões de acesso gratuito disponíveis à época da utilização** com o propósito de **testar a hipótese de pré-validação artificial dos itens gerados**. Após a utilização da ferramenta, os autores revisaram e editaram a apresentação dos resultados, sendo inteiramente responsáveis pelo conteúdo aqui apresentado. "

Recebido em: 24/02/2026
Parecer em: 08/04/2026
Aprovado em: 20/05/2026