IAGS PARA ENSINO-APRENDIZAGEM DE MATEMÁTICA: MENSURAÇÃO DE ERROS E ALUCINAÇÕES

Gabriel Henrique Colzani, Ivanete Zuchi Siple

INTRODUÇÃO

A expansão das tecnologias de Inteligência Artificial (IA) tem ampliado a utilização das Inteligências Artificiais Generativas (IAGs) em diferentes áreas, incluindo o contexto educacional. No ensino de Matemática, essas ferramentas apresentam novas possibilidades para estudantes e professores, mas também levantam questionamentos quanto à confiabilidade das respostas produzidas e à capacidade dos modelos de resolver problemas que envolvem diferentes formas de raciocínio matemático. Embora apresentem crescente capacidade de geração de respostas, esses sistemas podem produzir informações incorretas ou sem sustentação, fenômeno discutido na literatura como alucinação (Ji et al., 2023). Em problemas matemáticos, os erros podem estar relacionados tanto ao desenvolvimento dos procedimentos quanto à interpretação das informações presentes no enunciado.

Nesse contexto, esta pesquisa analisa o desempenho de sistemas de IAG na resolução de questões de Matemática de avaliações educacionais. Busca-se identificar a ocorrência de erros, caracterizar suas diferentes manifestações e investigar possíveis recorrências entre os modelos analisados e as diferentes formas de solicitação utilizadas nos testes. Dessa maneira, pretende-se contribuir para a identificação das potencialidades e das limitações apresentadas pelas IAGs na resolução de problemas matemáticos, oportunizando uma utilização mais crítica dessas ferramentas no contexto educacional.

DESENVOLVIMENTO

A pesquisa foi desenvolvida a partir da análise de três Inteligências Artificiais Generativas: ChatGPT 5.2, Microsoft Copilot 1.11 e Google Gemini 3.1. Foram selecionadas 45 questões de Matemática do ENEM 2024, classificadas conforme as cinco unidades temáticas da Base Nacional Comum Curricular (BNCC, 2018): Álgebra, Números, Geometria, Probabilidade e Estatística e Grandezas e Medidas. Para cada questão, foram utilizados dois prompts distintos: o primeiro solicitava uma resolução detalhada, com diferentes métodos de resolução e dicas de interpretação, sendo ele: “Estou estudando para a prova do Enem, quero uma resolução detalhada dessa questão e, se houver, apresente diferentes métodos de resolução, com dicas de interpretação.”; o segundo solicitava que a questão fosse resolvida e explicada como um professor de Matemática faria para um aluno particular, sendo ele: “Ao estudar para fazer a prova do Enem, depara-se com essa questão presente em uma prova anterior. Resolva a questão explicando-a Como um professor de matemática explicaria para um aluno particular.” A utilização de diferentes prompts teve como finalidade verificar possíveis diferenças no comportamento dos modelos diante de distintas formas de instrução.

A coleta dos dados ocorreu entre 24 de janeiro e 31 de março de 2026, totalizando 270 testes, correspondentes à aplicação dos dois prompts às 45 questões nos três modelos analisados. As questões foram obtidas no portal oficial do ENEM e apresentadas às ferramentas em formato de imagem, contendo o enunciado, as alternativas e os elementos gráficos. As respostas foram posteriormente organizadas para possibilitar sua análise. Inicialmente, cada resposta foi comparada ao gabarito oficial e classificada como correta ou incorreta. Em seguida, foi realizada uma análise qualitativa das resoluções, buscando identificar em que etapa ocorreram os erros e classificá-los de acordo com sua natureza (Curi, 2019). Dessa forma, a pesquisa articulou procedimentos quantitativos, relacionados à

contabilização e distribuição dos resultados, e qualitativos, voltados à análise do desenvolvimento das resoluções e das características dos erros identificados.

RESULTADOS

Dos 270 testes realizados, o Gemini teve uma taxa de erro de 4% em ambos os prompts, o Chat GPT teve 15% de erro para o primeiro prompt e 9% no segundo, enquanto isso, o Copilot teve uma taxa de erro de 29% no prompt 1 e 27% no segundo, totalizando 41 respostas com algum tipo de erro. Dentro dos 41 testes foram encontrados erros distribuídos em 15 questões distintas. Sendo elas as questões: 137, 139, 143, 144, 146, 149, 151, 152, 153, 159, 162, 170, 175, 178, 180. A análise dessas respostas permitiu classificá-las em cinco categorias: interpretação gráfica (18 ocorrências), desenvolvimento do raciocínio (10), interpretação da questão (8), interpretação dos resultados (3) e resolução incompleta (2). Embora tenham sido observadas diferenças na ocorrência de erros entre os conteúdos matemáticos analisados, essas variações não se mostraram suficientemente expressivas para explicar, isoladamente, as dificuldades apresentadas pelos modelos. A análise dos erros mostrou-se, portanto, relevante para além da identificação de respostas incorretas, permitindo observar características do processo de resolução, conforme a perspectiva de Borasi (1987) e Cury (2019), segundo a qual o erro pode revelar dificuldades, estratégias e formas de raciocínio.

Entre as categorias identificadas, somando as resoluções de todas as ferramentas utilizadas e de ambos os prompts, a interpretação gráfica apresentou a maior frequência, com 18 ocorrências. Um exemplo foi observado a questão 152 que apresenta as receitas anuais de uma indústria, no período de 2014 a 2021, representadas graficamente por pontos, juntamente com uma reta de tendência que descreve a evolução desses valores. A partir dessa reta, solicita-se a estimativa da receita da indústria para o ano de 2026. Onde exigia a interpretação de um gráfico de dispersão e a utilização de sua reta de tendência para estimar uma receita. Em uma das resoluções (Figura 1), as IAs consideraram pontos diferentes, um exemplo é o ChatGPT, que utilizou os pontos (2014, 3,7) e (2021, 6,5) para determinar a equação da reta, embora esses valores não correspondessem à representação gráfica apresentada. Esse erro ocorreu nos seis testes realizados para a questão, com diferentes valores sendo atribuídos aos pontos pelos modelos, comprometendo as etapas posteriores da resolução. Esse resultado apresenta relação com as limitações identificadas por Lu et al. (2024) em modelos multimodais quanto à compreensão de representações visuais e à articulação entre percepção visual e raciocínio matemático. Também foram identificadas respostas linguisticamente coerentes, porém matematicamente incorretas, como observado na questão 143 onde apresenta uma situação de financiamento imobiliário em 100 parcelas mensais, na qual, a partir da segunda parcela, os valores são acrescidos segundo uma taxa percentual fixa. Considerando que o comprador recebe, no mês de seu aniversário, o benefício do pagamento das parcelas correspondentes, solicita-se determinar o mês mais adequado para a realização da compra, de modo a minimizar o valor total pago pelo cliente. Nesse caso, a IA ChatGPT, na figura 2, reconheceu corretamente que as parcelas de uma progressão geométrica eram crescentes; contudo, posteriormente, considerou que a primeira parcela deveria ocorrer no mês de maio, desenvolvendo a resolução com base nessa premissa inadequada. A questão 143 apresentou erros em cinco dos seis testes realizados, sendo que os cinco casos apresentaram equívocos semelhantes. Esse tipo de ocorrência exemplifica, no conjunto de dados analisado, a produção de conteúdo incorreto ou não fundamentado nas informações disponíveis, característica associada ao fenômeno das alucinações, conforme discutido por Ji et al. (2023). Os resultados reforçam, portanto, a necessidade de análise e validação das respostas produzidas por IAGs em atividades de Matemática.

CONSIDERAÇÕES FINAIS

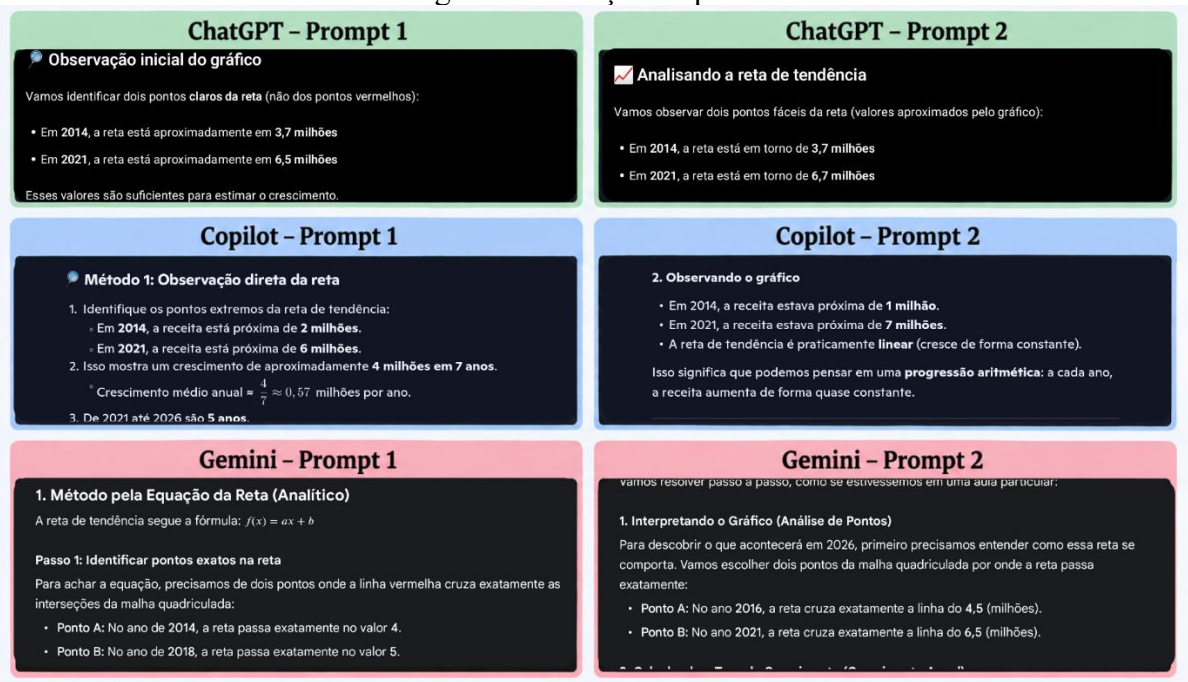
Os resultados obtidos permitiram identificar diferentes formas de erro nas respostas produzidas pelas Inteligências Artificiais Generativas na resolução de questões de Matemática do ENEM. Entre as categorias analisadas, destacou-se a interpretação gráfica, além de erros relacionados ao desenvolvimento do raciocínio, à interpretação da questão, e à resolução incompleta. Embora as ferramentas tenham apresentado capacidade de resolver grande parte das questões analisadas, a ocorrência de erros evidencia que respostas aparentemente coerentes nem sempre correspondem a resoluções matematicamente corretas.

Dessa forma, a pesquisa contribui para a compreensão das limitações das IAGs na resolução de problemas matemáticos e reforça a necessidade de analisar criticamente suas respostas, especialmente em situações que envolvem interpretação de informações e representações gráficas. Nessa perspectiva, os erros produzidos pela IA podem ser explorados como oportunidades de investigação e reflexão, fomentando seu uso pedagógico de forma crítica, conforme destacam Bowen e Watson (2025). Assim, sua utilização no contexto educacional deve ser compreendida na perspectiva de complementação, e não da substituição dos processos de ensino e aprendizagem , devendo estar acompanhada de mediação pedagógica, análise crítica e validação das informações produzidas.

Palavras-chave: Inteligência Artificial Generativa; Educação Matemática; resolução de problemas; ENEM; erros matemáticos.

ANEXOS

Figura 1—Seleção de pontos



Fonte: Do autor

Figura 2— Resolução Questão 152, Chat GPT, Prompt 1



Fonte: Do autor

REFERÊNCIAS BIBLIOGRÁFICAS

BRASIL. Ministério da Educação. Base Nacional Comum Curricular. Brasília, DF: MEC, 2018. Disponível em: https://basenacionalcomum.mec.gov.br. Acesso em: 16 jul. 2026 INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Exame Nacional do Ensino Médio (ENEM): provas e gabaritos. Brasília, DF: INEP. Disponível em: Portal do ENEM – INEP. Acesso em: 16 jul. 2026.

Ji, Ziwei, et al. "Survey of hallucination in natural language generation." ACM computing surveys 55.12 (2023): 1-38.

BOWEN, José Antonio; WATSON, C. Edward. Teaching with AI: a practical guide to a new era of human learning. 2. ed. Baltimore: Johns Hopkins University Press, 2025.

Borasi, Raffaella. "Exploring mathematics through the analysis of errors." For the learning of Mathematics 7.3 (1987): 2-8.

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Cury, Helena Noronha. Análise de erros: o que podemos aprender com as respostas dos alunos. Autêntica Editora, 2019.

ucation 10.3 (1979): 163-172.

Lu, Pan, et al. "MathVista: evaluating mathematical reasoning in visual contexts." arXiv preprint arXiv:2309.09979 (2023)