

INTELIGÊNCIA ARTIFICIAL GENERATIVA E MATEMÁTICA: UMA ANÁLISE COMPARATIVA DE DESEMPENHO NO ENEM 2024

Silvino Marques da Silva Junior ¹
Marli Teresinha Quartieri ²

RESUMO

A crescente aplicação de Inteligência Artificial Generativa em tarefas educacionais tem despertado interesse sobre sua eficácia na resolução de avaliações padronizadas. Este estudo analisa o desempenho comparativo dos modelos ChatGPT, Gemini e DeepSeek na prova de Matemática do ENEM 2024, utilizando o gabarito oficial para avaliar a precisão das respostas fornecidas por cada IA. A abordagem metodológica adotada foi a qualitativa, permitindo uma análise descritiva dos dados de desempenho das IA. Além da acurácia geral, o estudo investiga possíveis padrões de erro e a possível influência de questões que exigem interpretação de gráficos e tabelas. Os resultados obtidos mostram que os três modelos de IA tiveram um desempenho abaixo do esperado, com o modelo DeepSeek alcançando um resultado superior em relação aos outros modelos, com 51% de questões corretas, seguido pelo Gemini com 42,2% e pelo ChatGPT com 38% de acertos, demonstrando variações significativas entre os modelos na capacidade de interpretar e resolver problemas matemáticos. As IA demonstraram dificuldades em questões com figuras, gráficos e problemas de lógica, com padrões de erro que incluíam cálculos incorretos e interpretação inadequada dos dados. Os achados contribuem para a compreensão das limitações e potencialidades da IA generativa na resolução de problemas matemáticos em contextos acadêmicos e avaliativos, oferecendo subsídios para pesquisas futuras sobre a integração dessas tecnologias no ensino de matemática.

Palavras-chave: Inteligência Artificial Generativa, Matemática, ENEM, Avaliação Educacional.

¹ Doutorando do Curso de Pós-graduação em Ensino da Universidade do Vale do Taquari - UNIVATES, silvinomarques@ifpi.edu.br;

² Professora orientadora: Doutora, Programa de Pós-graduação em Ensino - UNIVATES, mtquartieri@univates.br;



INTRODUÇÃO

A emergência da Inteligência Artificial Generativa (IAGen) tem redefinido debates no campo educacional, especialmente diante do avanço da Web 4.0, caracterizada pela datificação e pelo uso intensivo de algoritmos para monitoramento e predição em tempo real (Santaella, 2020). A IAGen destaca-se por sua capacidade de criar conteúdos originais — textos, imagens, vídeos e códigos — a partir de grandes bases de dados, desafiando concepções tradicionais de criatividade e subjetividade (Santaella; Kaufman, 2024). Essa tecnologia é vista como a “quarta ferida narcísica” da humanidade, pois questiona a centralidade do humano na produção de conhecimento.

A distinção entre IA preditiva e generativa é fundamental para o debate ético e regulatório, já que cada abordagem demanda marcos específicos (Santaella, 2025). Além disso, a literatura aponta para a intensificação das desigualdades sociotécnicas, uma vez que grandes empresas de tecnologia controlam a cadeia cultural global, ampliando a brecha digital entre Norte e Sul (Santaella, 2020).

No contexto educacional, o uso de IAGen em avaliações como o ENEM suscita debates polarizados. Enquanto entusiastas veem potencial revolucionário, críticos alertam para limitações, riscos de “alucinações” e impactos na integridade acadêmica (Rudolph; Tan; Tan, 2023; Ferraro; Coelho, 2024). Estudos recentes mostram que, embora modelos como GPT-4o possam atingir acurácia próxima a 50% em exames de matemática, seu desempenho diminui em questões mais complexas ou que exigem interpretação visual (Yoon, 2024). Além disso, o uso indiscriminado de IAG pode levar a uma aprendizagem superficial, impactando negativamente o desempenho em avaliações que exigem raciocínio aprofundado (Wecks et al., 2024).

Por outro lado, há evidências de que a integração de feedback gerado por IA, quando combinada com a mediação humana, pode reduzir a carga cognitiva e favorecer estratégias de aprendizagem mais eficientes, especialmente em matemática (Cosentino et al., 2025). Isso sugere que o potencial da IA está em sua utilização como complemento, e não substituto, do ensino tradicional.

Diante desse cenário, este artigo propõe uma análise comparativa do desempenho dos modelos ChatGPT, Gemini e DeepSeek na prova de Matemática do ENEM 2024, utilizando o gabarito oficial como referência para avaliação da acurácia. A abordagem qualitativa adotada visa aprofundar a compreensão sobre o comportamento das IAs diante



de diferentes tipos de questões, com ênfase naquelas que envolvem interpretação visual e análise de dados. Os resultados obtidos buscam subsidiar discussões sobre o papel da IA generativa na educação, suas potencialidades e desafios, além de apontar caminhos para futuras pesquisas e aplicações.

REFERENCIAL TEÓRICO

As origens da Inteligência Artificial (IA) remontam aos anos 1950, quando programas como *General Problem Solver* e *ELIZA* demonstraram que máquinas poderiam executar tarefas de raciocínio e linguagem (Nah et al., 2023). Após um período de estagnação devido a limitações computacionais, a IA recuperou fôlego com o desenvolvimento das redes neurais profundas e do *deep learning*, que foi o principal motor para a revolução da IA nas últimas décadas, permitindo avanços expressivos em tarefas de linguagem, visão computacional e, especialmente, na criação de modelos generativos capazes de produzir conteúdos originais (Russell; Norvig, 2021). O impacto dessas técnicas é reconhecido como o fator-chave para a ascensão da IA generativa, que se diferencia por criar dados novos e realistas, e não apenas prever ou classificar informações com base em padrões existentes (Russell; Norvig, 2021).

A IAGen diferencia-se da IA preditiva ao criar conteúdos originais, extrapolando padrões dos dados de treinamento para gerar textos, imagens e códigos inéditos (Su & Yang, 2023). Modelos como o ChatGPT utilizam aprendizagem profunda para produzir respostas com aparência humana, enquanto outras ferramentas, como *Midjourney* e *DeepBrain*, expandem a atuação da IA para múltiplas modalidades (Nah et al., 2023). O desenvolvimento de modelos transmodais, como *text-to-image* (Qiao et al., 2022) e transformers audio-visuais (Wang et al., 2022), amplia ainda mais o escopo das aplicações, permitindo a criação de materiais educacionais multimodais e personalizados (Prasad, P.; Balse, R.; Balchandani, 2025).

No campo educacional, a IAGen é vista como uma inovação disruptiva, capaz de personalizar o ensino, automatizar feedbacks e apoiar tanto estudantes quanto docentes (Kasneci et al., 2023). Pesquisas destacam benefícios como a oferta de experiências de aprendizagem mais eficientes e adaptadas, além de facilitar a elaboração de materiais didáticos e avaliações (Pimentel; Carvalho, 2023). No entanto, a literatura também aponta desafios, como a necessidade de garantir a equidade no acesso, a transparência dos



algoritmos e a proteção de dados sensíveis dos estudantes (Ferraro; Coelho, 2024).

A integração da IAGen no ensino de matemática, por exemplo, pode potencializar a aprendizagem ao oferecer feedback imediato e tarefas personalizadas, mas exige que educadores desenvolvam competências específicas para utilizar essas ferramentas de forma crítica e responsável (Pimentel; Carvalho, 2023). O framework IDEE (Identificar objetivos, Determinar automação, Ética e Avaliação de eficácia) propõe que a adoção da IA educacional deve ser orientada por metas claras, níveis adequados de automação, consideração ética e avaliação contínua dos resultados (Su & Yang, 2023).

Do ponto de vista teórico, autores como Santaella e Kaufman (2024) e Rodrigues e Rodrigues (2023) ressaltam os paradoxos da IAGen: ela é simultaneamente aliada e ameaça, capaz e dependente, acessível e restritiva. Tais paradoxos exigem uma abordagem crítica, que reconheça tanto o potencial transformador quanto os riscos associados, como viés algorítmico, dependência excessiva e desafios à autenticidade das avaliações (Pimentel, Carvalho e Silveira, 2024).

Por fim, a literatura enfatiza que a adoção responsável da IAGen na educação requer políticas institucionais claras, formação docente contínua e o desenvolvimento de competências digitais e éticas por parte de todos os atores envolvidos (Ferraro; Coelho, 2024; Santaella, 2025). O equilíbrio entre inovação e reflexão crítica é fundamental para que a IAGen contribua de forma sustentável e equitativa para o avanço educacional.

METODOLOGIA

A pesquisa adotou uma abordagem qualitativa, com foco na análise descritiva do desempenho dos modelos ChatGPT, Gemini e DeepSeek na prova de Matemática do ENEM 2024. Inicialmente, foram selecionadas as versões mais recentes dos três modelos de IA, garantindo que todos operassem sob condições similares de acesso e parametrização. As questões da prova de Matemática do ENEM 2024 foram submetidas individualmente a cada modelo, utilizando o mesmo enunciado e alternativas apresentadas aos candidatos humanos.

As respostas geradas por cada IA foram coletadas e comparadas ao gabarito oficial do ENEM 2024, permitindo o cálculo da acurácia de cada modelo. Além da taxa de acertos, foram analisados os tipos de questões em que ocorreram erros, com especial atenção para aquelas que exigiam interpretação de gráficos, tabelas e figuras. Essa análise



permitiu identificar padrões de desempenho e limitações específicas de cada modelo, alinhando-se à literatura que aponta desafios na compreensão de elementos visuais por parte das IAs.

Para garantir a validade dos resultados, as respostas foram revisadas por professores de matemática, que avaliaram a pertinência das soluções apresentadas e classificaram os erros quanto à sua natureza (conceitual, interpretativa ou operacional). A análise qualitativa dos dados buscou compreender não apenas o desempenho quantitativo, mas também os fatores que influenciaram o sucesso ou fracasso das IAs em diferentes tipos de questões. Essa abordagem permitiu uma avaliação mais abrangente do potencial e das limitações dos modelos generativos no contexto educacional.

RESULTADOS E DISCUSSÃO

Os resultados revelaram que os três modelos de IA apresentaram desempenho inferior ao esperado na prova de Matemática do ENEM 2024, com variações significativas entre eles. O modelo DeepSeek obteve a maior taxa de acertos, com 51% das questões respondidas corretamente, seguido pelo Gemini com 42,2% e pelo ChatGPT com 38%. Esses dados corroboram estudos anteriores que apontam diferenças de desempenho entre modelos de LLMs, influenciadas por fatores como arquitetura, volume de dados de treinamento e estratégias de geração de respostas (Mittal *et al.*, 2023).

Quadro 1 - Desempenho Quantitativo dos Modelos

Modelo	Acertos (%)	Pontos Fortes	Principais Limitações
DeepSeek	51,0	Cálculos diretos, questões textuais	Interpretação visual, gráficos
Gemini	42,2	Cálculos, compreensão semântica	Gráficos, tabelas, figuras complexas
ChatGPT	38,0	Questões objetivas, texto	Elementos visuais, ambiguidade

A análise qualitativa evidenciou que as principais dificuldades dos modelos estavam concentradas em questões que exigiam interpretação de gráficos, tabelas e figuras, confirmando limitações já identificadas na literatura (Wang *et al.*, 2022). Em muitos casos, as IAs apresentaram respostas inconsistentes ou ignoraram informações visuais essenciais para a resolução correta do problema. Por outro lado, em questões



puramente textuais ou que exigiam apenas cálculos diretos, o desempenho foi superior, indicando que a compreensão semântica e a manipulação de dados numéricos são pontos fortes desses modelos (Kasneci *et al.*, 2023).

Outro aspecto relevante foi a identificação de padrões de erro recorrentes, como a tendência a fornecer respostas plausíveis, porém incorretas, quando confrontadas com informações ambíguas ou incompletas, fenômeno já documentado em avaliações internacionais (Davis, 2024). Conforme pode-se observar no Quadro 2 a seguir:

Quadro 2 - Comparativo dos principais padrões de erro

Tipo de Questão	Principais Dificuldades Observadas	Exemplos de Erro Comum
Figuras, gráficos, tabelas	Interpretação visual deficiente, omissão de dados	Ignorar eixos, legendas ou escalas
Problemas de lógica	Raciocínio abstrato e analogias insuficientes	Resposta plausível, mas incorreta
Cálculos matemáticos	Erros operacionais e de compreensão do enunciado	Cálculo incorreto, confusão de unidades

Esse comportamento reforça a necessidade de aprimoramento dos modelos, especialmente no que diz respeito à integração de capacidades multimodais e ao desenvolvimento de mecanismos de validação e supervisão (Qiao *et al.*, 2022). A comparação entre os modelos também sugere que abordagens como o RAG³ (*Retrieval-Augmented Generation*) podem contribuir para a melhoria do desempenho, ao fornecer contexto adicional e facilitar a interpretação de dados complexos (Nah *et al.*, 2023).

Em síntese, os resultados indicam que, embora as IAs generativas apresentem potencial para apoiar a resolução de questões matemáticas em avaliações padronizadas, ainda enfrentam desafios significativos, especialmente em tarefas que exigem interpretação visual e raciocínio avançado. O avanço dessas tecnologias depende do desenvolvimento de soluções que integrem diferentes modalidades de dados e promovam maior robustez e confiabilidade nas respostas geradas, alinhando-se às demandas do contexto educacional contemporâneo.

³ Refere-se a uma técnica usada em modelos de Inteligência Artificial generativa que combina geração de linguagem com mecanismos de busca contextual. Em vez de confiar apenas na base de dados usada no treinamento do modelo, o RAG recupera informações externas relevantes e as insere no contexto de geração, permitindo respostas mais atualizadas, precisas e fundamentadas.



CONSIDERAÇÕES FINAIS

A análise comparativa do desempenho dos modelos ChatGPT, Gemini e DeepSeek na prova de Matemática do ENEM 2024 permitiu evidenciar tanto os avanços quanto as limitações das Inteligências Artificiais Generativas (IAGen) em tarefas avaliativas complexas. Os resultados mostraram que, embora esses sistemas demonstrem competência em cálculos diretos e na resolução de questões textuais, ainda apresentam fragilidades significativas em itens que exigem interpretação visual, leitura de gráficos e raciocínio lógico. Tais limitações confirmam o que a literatura aponta sobre as barreiras cognitivas e multimodais enfrentadas por modelos de linguagem (Kasneci et al., 2023; Wang et al., 2022; Yoon et al., 2024).

Do ponto de vista pedagógico, os achados reforçam que a IA generativa deve ser compreendida não como substituta do professor, mas como instrumento complementar de mediação cognitiva, capaz de ampliar a personalização e o feedback em processos de aprendizagem (Pimentel e Carvalho, 2023). A incorporação dessas ferramentas exige, entretanto, formação docente continuada, políticas institucionais claras e desenvolvimento de competências éticas e digitais (Santaella, 2025).

Em síntese, a pesquisa indica que a integração da IAGen na educação precisa equilibrar inovação tecnológica e reflexão crítica. O futuro do ensino apoiado por IA dependerá da capacidade de promover ecossistemas híbridos — humanos e artificiais — baseados em transparência, equidade e responsabilidade. Ao mesmo tempo em que desafia concepções tradicionais de autoria e cognição, a IA generativa oferece oportunidades inéditas de criação de materiais educacionais multimodais e personalizados, alinhados às necessidades e contextos dos aprendizes.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.



REFERÊNCIAS

COSENTINO, Giulia et al. Hybrid teaching intelligence: Lessons learned from an embodied mathematics learning experience. *British Journal of Educational Technology*, 19 out. 2024. Disponível em: <https://doi.org/10.1111/bjet.13525>. Acesso em: 27 jan. 2025.

DAVIS, Ernest. Mathematics, word problems, common sense, and artificial intelligence. *Bulletin of the American Mathematical Society*, 15 fev. 2024. Disponível em: <https://doi.org/10.1090/bull/1828>. Acesso em: 30 abr. 2025.

FERRARO, Danielle Soares e. Silva Bicudo; COELHO, Márcia Azevedo. Há como deter o ChatGPT? Uma resenha da obra de Lúcia Santaella. *Galáxia* (São Paulo), v. 49, 2024. Disponível em: <https://doi.org/10.1590/1982-2553202465960>. Acesso em: 22 mai. 2025.

KASNECI, Enkelejda et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, v. 103, p. 102274, abr. 2023. Disponível em: <https://doi.org/10.1016/j.lindif.2023.102274>. Acesso em: 30 dez. 2024.

PIMENTEL, Mariano; CARVALHO, Felipe. ChatGPT: concepções epistêmico-didático-pedagógicas dos usos na educação. *SBC Horizontes*, v. 6, 2023.

PIMENTEL, Mariano; CARVALHO, Felipe; SILVEIRA, Victor Junger. IA Generativa pode ser coautora?. *Tríade: Comunicação, Cultura e Mídia*, v. 12, n. 25, p. e024012-e024012, 2024.

PRASAD, P.; BALSE, R.; BALCHANDANI, D. Exploring Multimodal Generative AI for Education through Co-design Workshops with Students. *CHI 2025: CHI Conference on Human Factors in Computing Systems*. ACM, 2025. p. 1-17. Disponível em: <https://doi.org/10.1145/3706598.3714146>. Acesso em: 16 abr. 2025.

QIAO, H., LIU, V., & CHILTON, L. (2022,). Initial images: Using image prompts to improve subject representation in multimodal AI generated art. *In Proceedings of the 14th Conference on Creativity and Cognition*, Venice, Italy, (pp. 15–28). <https://doi.org/10.1145/3527927.3532792>

RODRIGUES, Olira Saraiva; RODRIGUES, Karoline Santos. A inteligência artificial na educação: os desafios do ChatGPT. *Texto livre*, v. 16, p. e45997, 2023.

RUDOLPH, J., TAN, S., & TAN, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching*, 6(1). <https://doi.org/10.37074/jalt.2023.6.1.9>.

RUSSELL, Stuart J.; NORVIG, Peter. *Artificial Intelligence: A Modern Approach*, EBook, Global Edition. [S. l.]: Pearson Education, Limited, 2021. ISBN 9781292401171.



SANTAELLA, Lucia. *A pós-verdade é verdadeira ou falsa?*. Editora estação das letras e cores, 2020.

SANTAELLA, Lucia. Inteligência Artificial sob a Égide da Ética. *Revista Cronos*, v. 26, n. 1, p. 7-19, 2 mar. 2025. Disponível em: <https://doi.org/10.21680/1982-5560.2025v26n1id39309>. Acesso em: 13 abr. 2025.

SANTAELLA, Lucia; KAUFMAN, Dora. A Inteligência artificial generativa como quarta ferida narcísica do humano. *MATRIZES*, v. 18, n. 1, p. 37-53, 30 abr. 2024. Disponível em: <https://doi.org/10.11606/issn.1982-8160.v18i1p37-53>. Acesso em: 3 jun. 2025.

SU, J., & YANG, W. (2023). Unlocking the Power of ChatGPT: A Framework for Applying Generative AI in Education. *ECNU Review of Education*, 6(3), 355-366. <https://doi.org/10.1177/20965311231168423>

WANG, S., LI, L., DING, Y., & YU, X. (2022). One-shot talking face generation from single-speaker audio-visual correlation learning. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3), (pp. 2531–2539).

WECKS, Janik Ole et al. Generative AI Usage and Academic Performance. *SSRN Electronic Journal*, 2024. Disponível em: <https://doi.org/10.2139/ssrn.4812513>. Acesso em: 12 jan. 2025.

YOON, Hyunkyoung et al. Students' use of generative artificial intelligence for proving mathematical statements. *ZDM – Mathematics Education*, 26 ago. 2024. Disponível em: <https://doi.org/10.1007/s11858-024-01629-0>. Acesso em: 30 out. 2025.

