# LLMs São Bons Matemáticos? Avaliando o Desempenho em Resolução de Exercícios

**Igor Alberte R. Eleutério**[1,3], **Israel Efraim de Oliveira**[1,2], **Mirela T. Cazzolato**[3]

[1]Instituto das Cidades Inteligentes (ICI)
Curitiba, PR – Brasil

[2]Universidade Federal de Santa Catarina (UFSC)
Florianópolis, SC – Brasil

[3]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP) – São Carlos, SP – Brasil

igoralberte@gmail.com
israel.oliveira@posgrad.ufsc.br, mirela@icmc.usp.br

***Abstract.*** *This paper analyzes the use of two commercial LLM models (Google Gemini 1.5 Pro and OpenAI ChatGPT-4o) for solving high school mathematics exercises across five different topics, including Functions, Geometry, Combinatorics, and others. A total of 50 questions developed by the FGV (Getúlio Vargas Foundation) examination board were solved. Both models yielded similar results, with Gemini performing slightly better in terms of both reasoning and correct answer selection. In cases where ChatGPT made errors, its o3-pro version was able to provide the correct answers. The findings can inform decisions by individuals and organizations involved in mathematics teaching and learning regarding the use of this technology.*

***Resumo.*** *Este trabalho analisa o uso de dois modelos de LLMs comerciais (Google Gemini 2.5 Pro e OpenAI ChatGPT 4o) para resolução de exercícios de Matemática de Nível Médio sobre cinco diferentes tópicos que incluem Funções, Geometria, Análise Combinatória e outros. Ao todo, foram resolvidas 50 questões elaboradas pela banca FGV (Fundação Getúlio Vargas). Ambos os modelos tiveram resultados semelhantes, com o Gemini sendo ligeiramente melhor, tanto no raciocínio quanto na escolha da alternativa correta. Nas questões em que o ChatGPT apresentou erro, sua versão o3-pro foi capaz de acertá-las. Os resultados podem subsidiar decisões de pessoas e organizações envolvidas com ensino-aprendizagem de Matemática sobre o uso da tecnologia.*

## 1. Introdução

O processamento de linguagem natural, uma das principais frentes da inteligência artificial contemporânea, evoluiu de forma expressiva com o desenvolvimento recente dos Grandes Modelos de Linguagem (*Large Language Models* – LLMs) [Auffarth 2023]. Esses sistemas, baseados em técnicas de aprendizado profundo, mostram uma habilidade notável para gerar texto e interpretar informações complexas, abrindo caminho para aplicações em diversas áreas, tais como análise de dados [Miranda and Campelo 2024], segmentação de texto [Santos and Dorneles 2024] e educação [Marques and Morandini 2024].

Dado esse contexto, faz-se importante avaliar a precisão dessas ferramentas em diferentes áreas de ensino, o que ajuda a tomar decisões sobre adotá-las e de que forma. Neste trabalho, é avaliada a aplicação de LLMs para resolução de questões de Matemática de nível do Ensino Médio. Vários trabalhos já se dedicaram a avaliar a precisão dessas tecnologias nesse domínio [Collins et al. 2024, Gandolfi 2025, Satpute et al. 2024, Rodrigues et al. 2025], mas são escassos os trabalhos voltados para problemas escritos em Língua Portuguesa que exploram diversos assuntos (como Álgebra, Geometria, Combinatória) ao nível do Ensino Médio.

Esta pesquisa traz como questões de pesquisa:

- **Q1:** LLMs são ferramentas confiáveis para resolução de questões de Matemática, ao nível de Ensino Médio, escritas em português?
- **Q2:** LLMs são capazes de fornecer explicações corretas sobre a resolução de questões de Matemática ao nível de Ensino Médio?

A contribuição do trabalho, ao buscar responder às questões mencionadas, é de **fornecer subsídios a professores e demais envolvidos no processo educacional para decidirem sobre o uso de LLMs no contexto educacional, mais especificamente, no ensino-aprendizagem de Matemática**, levando em consideração a precisão dessas ferramentas e também os riscos envolvidos em seu uso.

O restante do artigo está organizado da seguinte forma: na Seção 2, são apresentados os Trabalhos Relacionados; na Seção 3, são especificados os Materiais e Métodos utilizados nos experimentos; na Seção 4, apresentam-se os Resultados e Discussão e, por fim, na Seção 5 são feitas as Considerações Finais.

## 2. Trabalhos Relacionados

LLMs têm apresentado bons resultados em vários cenários de aplicação, como para análise de dados [Miranda and Campelo 2024] e para anotações em *datasets* [Bencke et al. 2024]. Na educação não é diferente, e elas se mostram como ferramentas revolucionárias. Em [Wang et al. 2024] os autores trazem uma taxonomia de áreas de aplicação de LLMs no domínio educacional:

- **Suporte aos estudantes:** por exemplo, para resolução de questões e correção de erros;
- **Suporte aos professores:** em cenários como criação de materiais e correção automática;
- **Aprendizado adaptativo:** no caso de acompanhamento de aprendizagem e personalização de conteúdo.

Em todos os casos, é de suma importância avaliar a precisão das LLMs no cenário aplicado. Como exemplo, uma LLM que apresente uma baixa taxa de acertos em questões dificilmente poderá ser empregada como suporte aos estudantes.

### 2.1. LLMs para ensino-aprendizagem de Matemática

Alguns trabalhos da literatura focam na avaliação do uso de modelos de LLMs no processo de ensino-aprendizagem de Matemática em diferentes níveis, e dos impactos que essa tecnologia traz.

Em [Pardos and Bhandari 2024] os autores conduziram um estudo em que dicas eram fornecidas aos alunos durante a resolução de questões relacionadas à Matemática e à Estatística. Os autores concluem que dicas geradas pelo ChatGPT forneceram ganhos de aprendizagem aos alunos, de forma similar às dicas geradas por humanos, mas aquelas eram geradas de forma muito mais rápida do que estas.

Em [Liu et al. 2025] os autores desenvolveram um ambiente baseado em LLMs para dar suporte à resolução de questões de Matemática escritas em linguagem natural para alunos de ensino básico. Os autores perceberam que o uso dessa tecnologia trouxe dois principais benefícios: melhorou o desempenho dos estudantes na compreensão e resolução das atividades, e aumentou o interesse deles pela disciplina.

Com os exemplos descritos nesta subseção, é possível perceber que o uso de LLMs para o apoio ao processo de ensino-aprendizagem apresenta casos de sucesso.

### 2.2. LLMs para resolução de questões de Matemática

No trabalho [Collins et al. 2024] os autores avaliam o desempenho de LLMs para resolução de questões de Matemática de nível superior. Os autores desenvolvem uma ferramenta que permite aos alunos interagirem com o modelo durante a atividade, instruindo o modelo em caso de erros. Os autores recomendam sempre verificar as respostas fornecidas, por conta de erros cometidos. Na mesma linha, em [Gandolfi 2025] o autor avalia o ChatGPT, na versão GPT-4, para resolução e correção de questões de Cálculo. Nesse caso, embora o autor aponte problemas do modelo com aritmética, a precisão foi comparável a de humanos em alguns casos de correção de respostas. Já o trabalho [Satpute et al. 2024] aponta dificuldades de LLMs na resolução de questões de Matemática avançada. Os três estudos discutidos até aqui avaliam o desempenho desses modelos para questões escritas em inglês e para tópicos avançados da disciplina.

Um trabalho que busca preencher a lacuna de resolução de questões em português por LLMs é o de [Rodrigues et al. 2025]. Nesse caso, os autores fazem seus experimentos em diferentes tipos de questões (abertas, de múltipla escolha etc.), de Português e Matemática. Contudo, para as questões de Matemática, os únicos assuntos abordados são problemas que envolvem as quatro operações fundamentais (soma, subtração, divisão e multiplicação). Outros assuntos como Geometria, Análise Combinatória e Funções não são cobertos.

### 2.3. Riscos no uso de LLMs na educação

Algumas pesquisas se dedicam a avaliar e fornecer uma taxonomia dos riscos que as LLMs trazem à sociedade [Weidinger et al. 2022,Makridakis et al. 2023]. De forma geral, os riscos abrangem diversos aspectos como o meio-ambiente (uso intensivo de energia pelas LLMs), a sociedade (aumento de desigualdades sociais) e privacidade (vazamento de dados).

Outras pesquisas se dedicam a avaliar o risco de LLMs especificamente no âmbito educacional. No trabalho [Harvey et al. 2025] os autores conduzem uma análise a partir de entrevistas de fornecedores de tecnologias educacionais e de professores. Os autores levantam ameaças de diferentes tipos:

- **Ameaças técnicas:** conteúdo enviesado e tóxico gerado pelos modelos, violação de privacidade e alucinações;

- **Ameaças de interação:** desonestidade acadêmica devido ao uso indevido de LLMs para solucionar trabalhos e atividades;
- **Ameaças amplas:** inibição do aprendizado dos estudantes e de seu desenvolvimento social, aumento da carga de trabalho dos professores (que precisam se dedicar a aprender novas tecnologias a todo momento) e ampliação de desigualdades sistêmicas na educação (considerando que nem todos os estudantes têm acesso às tecnologias).

Um dos professores entrevistados na pesquisa levanta uma questão importante: "*Se um computador pode realizar a tarefa que você atribuiu, essa é a tarefa mais significativa?*".

Outro trabalho nesse sentido é o [Lehmann et al. 2025], em que os autores avaliam como LLMs impactam o aprendizado de alunos em um curso de graduação. Os autores concluem que o uso substitutivo de LLMs – aquele em que os alunos utilizam a tecnologia para substituir alguma tarefa de aprendizado, como para obter rapidamente a solução de um exercício – permite que os alunos ampliem o número de tópicos estudados, mas faz com que a compreensão de cada tópico seja superficial. Os autores recomendam que, preferencialmente, as LLMs devem ser usadas para complementar as atividades de aprendizagem ao invés de substituí-las.

### 2.4. Abordagem desta pesquisa

Nesta pesquisa, seguindo a taxonomia proposta por [Wang et al. 2024], serão avaliadas duas LLMs comerciais quanto à sua capacidade de resolver e explicar soluções corretamente para problemas de Matemática de Ensino Médio, tendo em vista sua possível aplicação em caso de **Suporte aos Estudantes**, para resolução de questões, e de **Aprendizado Adaptativo**, para suporte personalizado à aprendizagem. Essa análise se justifica pela escassez de trabalhos que avaliam essa tecnologia no domínio de questões de Matemática escritas em Língua Portuguesa. Mais especificamente, questões sobre diversos assuntos da Matemática do Ensino Médio serão abordadas.

Esta pesquisa busca apontar a precisão dessas tecnologias nesse domínio e lançar luz para caminhos possíveis para professores e desenvolvedores de tecnologias educacionais. Os riscos levantados nos trabalhos relacionados também são importantes por permitirem uma discussão crítica sobre o uso dessas ferramentas na educação.

## 3. Materiais e Métodos

O fluxo de execução dos experimentos deste trabalho está esquematizado na Figura 1. Nos passos 1 e 2, as questões selecionadas são enviadas às LLMs **Google Gemini 2.5 Pro** e **ChatGPT 4o**. Nos casos em que essa versão do ChatGPT apresenta erro na resposta, a questão também é resolvida utilizando-se a versão **GPT o3-pro** (passo 3), que é focado em reflexão e conseguiu resolver algumas das questões erradas pelo modelo 4o corretamente.

Para cada resolução, **avaliou-se tanto a Alternativa escolhida pela LLM quanto o Raciocínio.** Isso se deve ao fato de que, para uso das LLMs como ferramenta educacional, é fundamental que o modelo consiga descrever corretamente os passos seguidos para se chegar à resposta e não apenas dar a alternativa correta. Em um cenário de uso

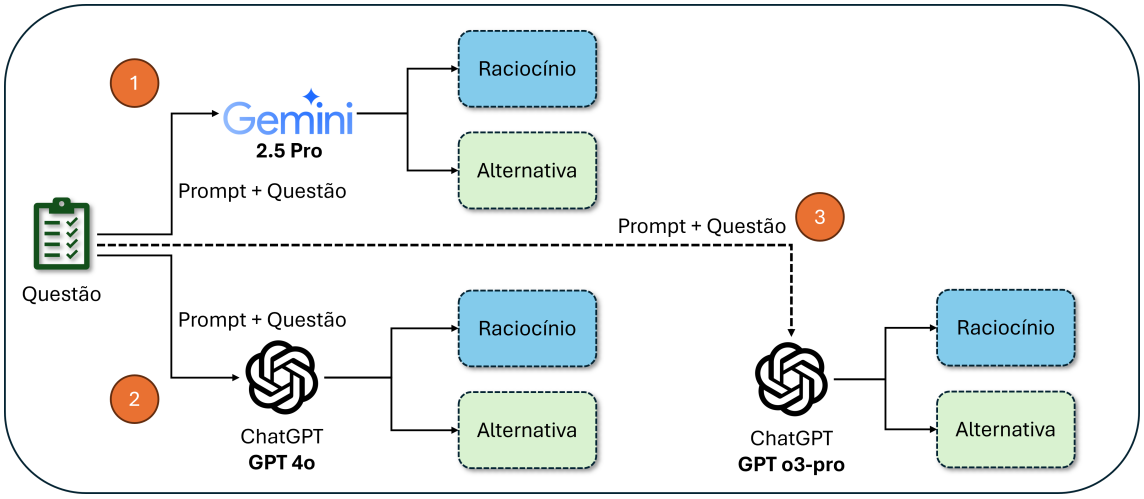


**Figura 1. Fluxo de execução dos experimentos.**

em que a LLM é utilizada como tutor, essa capacidade é importante. Para a avaliação da corretude, foi realizada uma avaliação manual de cada resposta. Assim, Raciocínios e Alternativas foram considerados como corretos ou errados separadamente, porque a LLM poderia acertar a Alternativa, mas errar o Raciocínio (ou vice-versa).

Como essas ferramentas sofrem alterações e evoluções constantes, registra-se que os testes foram feitos entre os dias 18/06/2025 e 29/06/2025.

### 3.1. Questões selecionadas

Foram selecionadas 50 questões de tópicos ligados à Matemática do Ensino Médio. Essas questões foram extraídas de provas objetivas de concursos públicos desenvolvidas pela banca FGV (Fundação Getúlio Vargas)[1], instituição bem reconhecida na organização e condução de certames públicos no país. Foram 10 questões de múltipla escolha de cada um dos tópicos a seguir:

- **Raciocínio Lógico:** envolve a manipulação lógica de sentenças e símbolos utilizando regras como equivalências e negações de proposições;
- **Álgebra e Funções:** envolve a organização e manipulação algébrica, além de questões sobre funções de primeiro e segundo grau e funções exponenciais;
- **Razão e Proporção:** envolve a manipulação de grandezas que apresentam proporcionalidade direta ou inversa;
- **Geometria:** questões que trabalham conceitos da Geometria, como ângulos, perímetros, áreas e volumes;
- **Análise Combinatória e Probabilidade:** exercícios que envolvem o agrupamento de objetos utilizando conceitos como Combinação e Arranjo.

Foram utilizadas apenas questões sem imagens. Ou seja, todas as questões foram descritas apenas por meio de textos, mesmo quando havia equações e incógnitas envolvidas.

A Tabela 1 ilustra exemplos de questões selecionadas para os experimentos. Todas as questões são de múltipla escolha (possui alternativas, com apenas uma correta) e o gabarito é fornecido pela banca FGV para os certames em que a questão foi utilizada.

[1]Provas da FGV: https://conhecimento.fgv.br/concursos

**Tabela 1. Exemplos de questões de concursos públicos criadas pela banca FGV e selecionadas para os experimentos.**

| **Raciocínio Lógico** |
|---|
| Considere verdadeira a seguinte declaração:<br>***"Se eu acordo tarde, não faço desjejum."***<br>É correto concluir que<br>Alternativas<br>A) se eu não acordo tarde, faço desjejum.<br>B) se eu não acordo tarde, também não faço desjejum.<br>C) se eu faço o desjejum, então acordei tarde.<br>D) se eu faço o desjejum, então não acordei tarde. |
| **Álgebra e Funções** |
| Sejam A e B as raízes da equação $x^2 - 7x + 4 = 0$.<br>O valor de $A^2 + B^2$ é:<br>Alternativas<br>A) 49; B) 41; C) 36; D) 28; E) 11. |
| **Razão e Proporção** |
| Uma grandeza Y é diretamente proporcional a uma grandeza X e inversamente proporcional ao quadrado de uma grandeza Z. As três grandezas só assumem valores positivos.<br>Sabe-se que quando $X = 3$ e $Z = 2$, tem-se $Y = 0,9$.<br>Assim, quando $Y = 2$ e $X = 15$, o valor de $Z$ é<br>Alternativas<br>A) 2. B) 3. C) 6. D) 9. E) 12. |
| **Geometria** |
| Os pontos $A(2,0)$ e $B(0,2)$, identificados no sistema de coordenadas cartesianas, são vértices consecutivos do quadrado ABCD.<br>Se os demais vértices desse quadrado estão no 1º quadrante, um desses vértices pode ser identificado como um ponto de coordenadas<br>Alternativas<br>A) (2,2). B) (4,4). C) (4,0). D) (2,4). E) (0,4). |
| **Análise Combinatória e Probabilidade** |
| Em um pote há cinco balas sendo duas de menta e três de morango, todas de mesmo aspecto e tamanho. Joãozinho retira, ao acaso, duas balas desse pote.<br>A probabilidade de que Joãozinho tenha retirado duas balas do mesmo sabor é:<br>Alternativas<br>A) 20%. B) 30%. C) 40%. D) 50%. E) 60%. |

## 3.2. Prompts utilizados

Os *prompts* utilizados, mostrados nas Tabelas 2 e 3, foram simples e traziam apenas a instrução de não serem utilizados símbolos que não pudessem ser copiados em um documento Microsoft Office Word. Mesmo com o *prompt*, as ferramentas muitas vezes trouxeram símbolos especiais que não foram adequadamente transpostos para o documento do Word. Após a instrução inicial, foi copiada a questão e suas alternativas logo abaixo.

**Tabela 2. Modelo de prompt utilizado para o Gemini.**

| Prompt para o Gemini | |
|---|---|
| 1 | Responda esta questão e dê como resposta um texto que possa ser copiado e colado no Word. |
| 2 | ***<Enunciado da questão>*** |
| 3 | ***<Alternativas da questão>*** |

**Tabela 3. Modelo de prompt utilizado para o ChatGPT.**

| Prompt para o ChatGPT | |
|---|---|
| 1 | Responda esta questão e dê como resposta um texto plano, sem símbolos especiais nas fórmulas, para que possa ser copiado e colado no Word. |
| 2 | ***<Enunciado da questão>*** |
| 3 | ***<Alternativas da questão>*** |

### 3.3. Experimento complementar – questões sem alternativas

Posteriormente, foi executado um experimento similar ao primeiro, mas apenas com os enunciados das questões, sem as alternativas, mantendo-se os prompts. Buscou-se averiguar se a presença das alternativas influenciava o desempenho dos modelos de LLMs. O conjunto de questões testado foi o mesmo do primeiro experimento, excetuando-se as questões cujas resoluções estavam muito dependentes das alternativas. Para este experimento, foram utilizadas 39 questões.

## 4. Resultados e Discussão

As questões e respostas dadas pelos modelos, bem como a análise do tamanho das respostas, estão disponíveis para consulta em um repositório online[2]. A Tabela 4 resume os resultados dos experimentos (questões com alternativas) e mostra que ambos os modelos tiveram desempenhos ótimos nas questões propostas, tanto para escolher a Alternativa correta quanto para especificar o Raciocínio ao buscar a resposta.

Com exceção de uma questão em que o ChatGPT 4o errou a Alternativa e o Raciocínio, e de outra em que ele errou o Raciocínio, mas escolheu a Alternativa correta, todas as outras questões foram acertadas pelos dois modelos. Para as duas questões erradas, o modelo ChatGPT o3-pro foi solicitado a respondê-las e obteve sucesso. Esse resultado indica que, no estágio atual de desenvolvimento, e para questões de Nível Médio de múltipla escolha, ambas LLMs geram respostas muito satisfatórias.

### 4.1. Detalhamento das respostas

Além de avaliar a corretude das Alternativas fornecidas como resposta, é importante avaliar o Raciocínio, pois uma ferramenta que se proponha a dar suporte ao ensino-aprendizagem precisa ser precisa ao fornecer dicas ou soluções de exercícios.

Nesse quesito, **o Gemini apresentou respostas mais detalhadas**. Em média, o tamanho da resposta do Gemini foi de 319 palavras, contra 198 palavras do ChatGPT.

[2]Materiais suplementares: https://github.com/igoralberte/sbbd2025-llms-matematica

**Tabela 4. Desempenho Comparativo dos Motores Gemini (2.5 pro) e ChatGPT (4o). São apresentados o número de acertos de Alternativa e Raciocínio, por Categoria, e a Média Geral.**

| Categoria | Motor | Número de acertos | |
|---|---|---|---|
| | | Alternativa | Raciocínio |
| Algebra e Funções | Gemini | 10 | 10 |
| | ChatGPT | 10 | 10 |
| Análise Combinatória e Probabilidade | Gemini | 10 | 10 |
| | ChatGPT | 10 | 10 |
| Geometria | Gemini | 10 | 10 |
| | ChatGPT | 9 | 8 |
| Raciocínio Lógico | Gemini | 10 | 10 |
| | ChatGPT | 10 | 10 |
| Razão, Proporção e Regra de Três | Gemini | 10 | 10 |
| | ChatGPT | 10 | 10 |
| **Média Geral** | **Gemini** | **10.0** | **10.0** |
| | **ChatGPT** | **9.8** | **9.6** |

Nessa análise, foram desconsideradas a introdução e a conclusão das respostas, que não trazem informações úteis para as questões, bem como a cópia do enunciado que o Gemini sempre traz nas respostas.

### 4.2. Questões de Geometria respondidas incorretamente pelo ChatGPT

Os únicos erros do GPT 4o estavam relacionados à interpretação dos problemas em duas questões de Geometria, ilustradas na Figura 2. Na Questão (a) [Questão 1 da lista de Geometria], que tratava de esferas e uma mesa, o principal equívoco foi não considerar o diâmetro da esfera que estava no chão ao calcular a altura da mesa, utilizando erroneamente apenas o raio, resultando na representação incorreta das distâncias. Já na Questão (b) [Questão 3 da lista de Geometria], que envolvia um triângulo retângulo, o modelo falhou ao não reconhecer que o maior ângulo de um triângulo retângulo é sempre $90°$, afirmando que o maior dos ângulos agudos era o maior ângulo do triângulo.

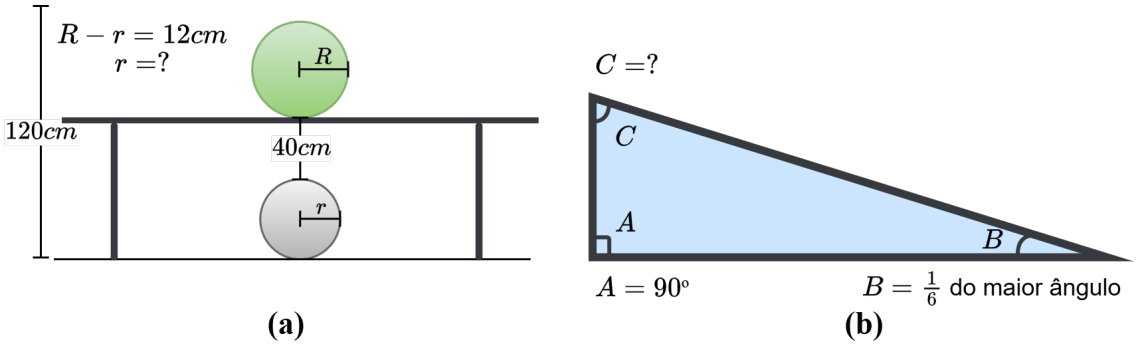


**Figura 2. Questões de Geometria: (a) Questão 1, sobre o valor do raio da menor esfera; (b) Questão 3, sobre o maior dos ângulos agudos.**

### 4.3. Experimento complementar – questões sem alternativas

Neste experimento, cujos resultados também estão no repositório do artigo, o desempenho dos modelos foi similar ao desempenho obtido no experimento anterior.

A única diferença foi que o ChatGPT cometeu um equívoco em outra questão de Geometria (Questão 2), além dos equívocos cometidos no experimento anterior que voltaram a se repetir. Dessa vez, nem o modelo o3-pro foi capaz de acertar. A questão solicitava o ângulo entre dois segmentos de reta e fornecia o ângulo entre cada segmento de reta e o norte geográfico.

Entretanto, o Gemini manteve sua precisão e, nas demais questões, o ChatGPT também continuou acertando. Isso indica que o desempenho dos modelos em questões de Matemática de Ensino Médio, com e sem alternativas, é muito bom.

### 4.4. Reflexões levantadas

Com a precisão apresentada pelos modelos, algumas questões relacionadas a riscos no âmbito educacional precisam ser levantadas. A **forma de avaliar os alunos**, com o advento das LLMs, precisa ser repensada. Em um cenário em que atividades propostas para o aluno praticar os assuntos aprendidos podem ser completamente respondidas por máquinas, acende-se um alerta sobre a eficiência desses métodos de avaliação. Outra questão importante é a necessidade de se **discutir o que de fato é importante que os alunos aprendam e quais habilidades são relevantes, para a sua vida profissional e pessoal**, que precisam ser abordadas na escola. Esse aspecto é importante sobretudo no âmbito da Matemática, em uma realidade em que as atividades propostas podem ser respondidas completamente com o uso de tecnologias.

## 5. Considerações Finais

Para ambas as perguntas inicialmente propostas nesta pesquisa, a saber:

- **Q1:** LLMs são ferramentas confiáveis para resolução de questões de Matemática, ao nível de Ensino Médio, escritas em português?
- **Q2:** LLMs são capazes de fornecer explicações corretas sobre a resolução de questões de Matemática ao nível de Ensino Médio?

Com base na observação experimental realizada, pode-se responder que “Sim”. Nos experimentos realizados, as LLMs comerciais foram capazes de acertar praticamente todas as questões (o Gemini 2.5 Pro conseguiu 100% de acerto), tanto em termos de Raciocínio quanto em termos da Alternativa escolhida. Logo, ferramentas de apoio ao aprendizado e de geração de materiais didáticos podem se aproveitar desse tipo de tecnologia.

Entretanto, esses resultados também sugerem a necessidade de repensar os métodos de ensino-aprendizagem atuais, sobretudo no âmbito da Matemática. Em uma realidade em que máquinas são capazes de solucionar os desafios propostos pela escola, talvez seja preciso rediscutir o que, de fato, deve ser aprendido.

Como limitações, é salutar destacar que esta pesquisa avaliou apenas questões de Nível Médio. Outra limitação é sobre o baixo número de questões avaliadas (50). Investigar um número maior de questões, e que também envolvam outros assuntos e outros tipos

de exercícios, além do desempenho das LLMs em diferentes disciplinas, são oportunidades de trabalhos futuros promissores para uma avaliação mais ampla do poder dessas tecnologias na Educação.

## Agradecimentos

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código de Financiamento 001, da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP – processos #2024/13328-9 e #2024/15430-5), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), da Universidade de São Paulo (PRPI 1032, processo #207), da Fundação de Ensino e Engenharia de Santa Catarina, nº contrato FEESC/PETROBRAS/4600671602 e do Instituto das Cidades Inteligentes (ICI - Curitiba/PR).

## Referências

Auffarth, B. (2023). *Generative AI with LangChain*. Packt Publishing, Birmingham, England.

Bencke, L., Paula, F., dos Santos, B., and Moreira, V. P. (2024). Can we trust LLMs as relevance judges? In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 600–612, Porto Alegre, RS, Brasil. SBC. DOI: http://dx.doi.org/10.5753/sbbd.2024.243130.

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., and Jamnik, M. (2024). Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121. DOI: 10.1073/pnas.2318124121.

Gandolfi, A. (2025). GPT-4 in Education: Evaluating Aptness, Reliability, and Loss of Coherence in Solving Calculus Problems and Grading Submissions. *International Journal of Artificial Intelligence in Education*, 35:367–397. DOI: http://dx.doi.org/10.1007/s40593-024-00403-3.

Harvey, E., Koenecke, A., and Kizilcec, R. F. (2025). "Don't Forget the Teachers": Towards an Educator-Centered Understanding of Harms from Large Language Models in Education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery. DOI: http://dx.doi.org/10.1145/3706598.3713210.

Lehmann, M., Cornelius, P. B., and Sting, F. J. (2025). AI Meets the Classroom: When Do Large Language Models Harm Learning? SSRN. DOI: http://dx.doi.org/10.2139/ssrn.4941259.

Liu, J., Sun, D., Sun, J., Wang, J., and Yu, P. L. H. (2025). Designing a generative AI enabled learning environment for mathematics word problem solving in primary schools: Learning performance, attitudes and interaction. *Computers and Education: Artificial Intelligence*, 9:100438. DOI: https://doi.org/10.1016/j.caeai.2025.100438.

Makridakis, S., Petropoulos, F., and Kang, Y. (2023). Large language models: Their success and impact. *Forecasting*, 5(3):536–549. DOI: http://dx.doi.org/10.3390/forecast5030030.

Marques, D. and Morandini, M. (2024). Uso do ChatGPT no Contexto Educacional: Uma Revisão Sistemática da Literatura. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação (SBIE 2024)*, SBIE 2024, page 1784–1795. Sociedade Brasileira de Computação - SBC. DOI: http://dx.doi.org/10.5753/sbie.2024.242535.

Miranda, B. and Campelo, C. E. C. (2024). How effective is an LLM-based Data Analysis Automation Tool? A Case Study with ChatGPT's Data Analyst. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 287–299, Porto Alegre, RS, Brasil. SBC. DOI: http://dx.doi.org/10.5753/sbbd.2024.240841.

Pardos, Z. A. and Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLOS ONE*, 19(5):1–18. DOI: http://dx.doi.org/10.1371/journal.pone.0304013.

Rodrigues, L., Xavier, C., Costa, N., Batista, H., Silva, L. F. B., Chaleghi de Melo, W., Gasevic, D., and Ferreira Mello, R. (2025). LLMs Performance in Answering Educational Questions in Brazilian Portuguese: A Preliminary Analysis on LLMs Potential to Support Diverse Educational Needs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, page 865–871, New York, NY, USA. Association for Computing Machinery. DOI: http://dx.doi.org/10.1145/3706468.3706515.

Santos, V. S. and Dorneles, C. F. (2024). Unveiling the Segmentation Power of LLMs: Zero-Shot Invoice Item Description Analysis. In *Anais do XXXIX Simpósio Brasileiro de Banco de Dados (SBBD 2024)*, SBBD 2024, page 549–561. Sociedade Brasileira de Computação - SBC. DOI: http://dx.doi.org/10.5753/sbbd.2024.240820.

Satpute, A., Gießing, N., Greiner-Petter, A., Schubotz, M., Teschke, O., Aizawa, A., and Gipp, B. (2024). Can LLMs Master Math? Investigating Large Language Models on Math Stack Exchange. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2316–2320, New York, NY, USA. Association for Computing Machinery. DOI: http://dx.doi.org/10.1145/3626772.3657945.

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., and Wen, Q. (2024). Large Language Models for Education: A Survey and Outlook. URL: https://arxiv.org/abs/2403.18105.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., and Gabriel, I. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery. DOI: http://dx.doi.org/10.1145/3531146.3533088.