PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

CURSO DE ENGENHARIA DE SOFTWARE

BERNARDO DE RESENDE FLÁVIO JÚNIOR
JOÃO MARCELO MIGUEL DINIZ

ANÁLISE DA PRECISÃO DE IAS GENERATIVAS NA RESOLUÇÃO DE EQUAÇÕES MATEMÁTICAS

Belo Horizonte - MG

2025

BERNARDO DE RESENDE FLÁVIO JÚNIOR
JOÃO MARCELO MIGUEL DINIZ

ANÁLISE DA PRECISÃO DE IAS GENERATIVAS NA RESOLUÇÃO DE EQUAÇÕES MATEMÁTICAS

Trabalho apresentado à disciplina de Introdução à Pesquisa em Informática do Curso de Engenharia de Software da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para avaliação.

Orientador: Danilo de Quadros Maia Filho

Belo Horizonte - MG

2025

RESUMO

Este trabalho investiga o desempenho de sistemas de inteligência artificial generativa na resolução de equações matemáticas complexas, com foco em integrais de diferentes níveis de dificuldade. A pesquisa justifica-se pela crescente aplicação de IAs generativas, como ChatGPT-4o, Claude Sonnet 4, Gemini 2.5 Pro, Microsoft Copilot, Grok 3 e DeepSeek R1, em contextos educacionais e computacionais, apesar de suas limitações em problemas matemáticos avançados. O objetivo geral é avaliar a precisão desses sistemas em resolver 21 integrais selecionadas. A metodologia envolve uma análise quantitativa e qualitativa dos resultados obtidos a partir de testes experimentais com cinco sistemas de IA, utilizando o cálculo integral como fundamentação teórica.

Palavras-chave: Inteligência artificial, IAs generativas, resolução de equações, cálculo integral, computational power.

1 Introdução

A inteligência artificial (IA) generativa tem emergido como um paradigma transformador em diversas áreas, incluindo a resolução de problemas matemáticos complexos. Modelos como GPT-4o, Claude Sonnet 4, Gemini 2.5 Pro, Microsoft Copilot, Grok 3 e DeepSeek R1, baseados em redes neurais artificiais e large language models, são amplamente utilizados em contextos educacionais e computacionais, transformando processos de ensino e aprendizagem. Contudo, estudos recentes, como o de Servick (2024), indicam que esses sistemas apresentam taxas de acerto inferiores a 2% em problemas matemáticos avançados, destacando a necessidade de avaliar sua eficácia em tarefas específicas, como a resolução de equações matemáticas.

A questão-problema desta pesquisa é: qual é a precisão de sistemas de IA generativa na resolução de integrais de diferentes níveis de complexidade, e quais são suas limitações em detectar erros procedimentais? Resolver esse problema é relevante, pois a aplicação de IAs no ensino e na prática matemática pode melhorar a eficiência do aprendizado, mas exige uma compreensão clara de suas capacidades e falhas. A justificativa para este estudo reside na importância de desenvolver ferramentas de IA mais robustas para aplicações educacionais, especialmente em disciplinas como cálculo diferencial e integral, que são fundamentais para a formação em ciências exatas.

A metodologia adotada consiste em uma avaliação comparativa de seis sistemas de IA (GPT-4o, Claude Sonnet 4, Gemini 2.5 Pro, Microsoft Copilot, Grok 3 e DeepSeek R1), utilizando um protocolo experimental com 21 integrais de diferentes níveis de complexidade, fundamentado em Stewart (2022). A análise será quanti-qualitativa, examinando a taxa de acertos e os tipos de erros cometidos pelos sistemas. Este trabalho contribuirá para o entendimento das limitações atuais das IAs generativas e para o aprimoramento de sua aplicação em contextos acadêmicos.

Este trabalho está organizado da seguinte forma. A Seção 2 apresenta os objetivos geral e específicos da pesquisa. A Seção 3 apresenta o referencial teórico, abordando conceitos de inteligência artificial, redes neurais, large language models e pensamento computacional, além de trabalhos relacionados. A Seção 4 descreve os procedimentos metodológicos, detalhando o protocolo experimental e os sistemas avaliados. As seções subsequentes abordarão os resultados, a discussão e as conclusões do estudo.

2 Objetivos

2.1 Objetivo Geral

Avaliar a precisão de sistemas de inteligência artificial generativa na resolução de equações matemáticas, com foco em 21 integrais de diferentes níveis de complexidade, identificando suas capacidades e limitações.

2.2 Objetivos Específicos

1. Comparar o desempenho de seis sistemas de IA (ChatGPT-4o, Claude Sonnet 4, Gemini 2.5 Pro, Grok 3, Microsoft Copilot e DeepSeek R1) na resolução de integrais.

2. Identificar os tipos de erros procedimentais cometidos pelos sistemas de IA durante a resolução das equações.

3. Propor recomendações para o aprimoramento de IAs generativas em aplicações educacionais de matemática, com base nos resultados da análise.

3 Revisão Bibliográfica

Esta seção está organizada em duas partes: Fundamentação Teórica e Trabalhos Relacionados. A primeira aborda conceitos essenciais sobre o uso da Inteligência Artificial para resolução de operações matemáticas complexas, como integrais. A segunda apresenta e analisa estudos recentes relacionados ao tema, que fundamentam a proposta deste trabalho.

3.1 Fundamentação Teórica

O uso de IAs para resolver problemas matemáticos complexos, como integrais, destaca-se como um avanço significativo na computação e na educação. Modelos contemporâneos, como os Large Language Models (LLMs), baseiam-se em arquiteturas do tipo Transformer e aprendem padrões estatísticos a partir de grandes volumes de dados, permitindo-lhes gerar respostas contextualizadas e resolver operações matemáticas de maneira autônoma [3].

No campo da resolução de integrais, as LLMs precisam identificar a estrutura da expressão ou modelo matemático e aplicar a técnica adequada, como substituição, integração por partes, integração trigonométrica, entre outros métodos de resolução. LLMs avançados são capazes de simular esse processo ao aprender padrões de manipulação algébrica, gerando inclusive códigos em Python para resolver tais operações, como demonstrado por [9].

Além das redes neurais voltadas à linguagem natural, existem abordagens específicas para a matemática simbólica, que envolvem a manipulação algébrica direta. A combinação de análise simbólica com programação automática, denominada *program synthesis*, permite que modelos atinjam desempenhos comparáveis aos humanos na resolução de problemas típicos do cálculo universitário [9, 3].

Paralelamente, sistemas puramente simbólicos, como os sistemas de álgebra computacional (CAS), garantem precisão, mas não possuem a capacidade explicativa dos LLMs. Assim, modelos híbridos emergem como alternativa promissora, integrando a clareza das explicações com a precisão dos cálculos. Esse panorama justifica a importância de estudos e análises aprofundadas que avaliem diferentes tipos de IA quanto à sua eficácia na resolução de integrais.

3.2 Trabalhos Relacionados

Diversos estudos recentes exploram o uso da IA na resolução de problemas matemáticos, especialmente integrais, sendo fundamentais para a estruturação deste trabalho.

3.2.1 Drori et al. (2021)

[9] propõem um método que combina *few-shot learning* com *program synthesis* para resolver problemas matemáticos de nível universitário. Utilizando o modelo Codex, da OpenAI, o sistema gera scripts em Python utilizando a biblioteca Sympy, alcançando 81% de precisão na resolução de questões de cálculo diferencial e integral. Este estudo evidencia como a integração de modelos de linguagem e motores simbólicos pode automatizar e otimizar a resolução de integrais, com explicações detalhadas e um passo a passo bem elaborado. Tal cenário motiva a análise proposta neste trabalho, considerando aspectos como precisão, velocidade de resposta e clareza das soluções geradas por diferentes modelos.

3.2.2 Zhong et al. (2023)

[3] apresentam o MathVista, um conjunto de dados destinado a avaliar a capacidade de modelos de linguagem em realizar raciocínio matemático, incluindo operações complexas de cálculo. O estudo avalia modelos como LLaMA e GPT-3, utilizando métricas de precisão e explicação de cada LLM. A proposta de [3] é referência para a definição dos *benchmarks* utilizados nesta pesquisa, possibilitando a comparação objetiva entre modelos quanto à sua habilidade de resolver integrais e outras operações complexas.

3.2.3 Pereira et al. (2024)

[5] relatam uma experiência prática de integração de IA no ensino de cálculo, comparando o desempenho de turmas que utilizaram uma plataforma baseada em IA com aquelas que seguiram métodos tradicionais. Os resultados indicaram melhora significativa na taxa de acertos em integrais e maior engajamento dos estudantes. Esse estudo destaca a aplicabilidade e eficiência da IA no ensino de cálculo e fornece subsídios metodológicos para a análise de desempenho dos modelos de IA neste trabalho, especialmente quando se analisa a precisão e a receptividade das soluções geradas.

3.2.4 Souza et al. (2024)

[6] realizam uma análise comparativa de ferramentas de IA, como Photomath, GeoGebra e ChatGPT, aplicadas ao ensino da matemática. O estudo aborda funcionalidades, limitações e impactos no processo de aprendizagem, incluindo operações de cálculo avançado. A contribuição deste trabalho está na identificação de diferentes abordagens tecnológicas disponíveis para a resolução de integrais, permitindo a contextualização e fundamentação da escolha dos modelos comparados nesta pesquisa.

3.2.5 Azevedo (2024)

[7] avalia o uso do ChatGPT na resolução de problemas matemáticos aplicados, como escalas e cálculos de área, abordando também a geração de integrais simples por meio de *prompts* em linguagem natural. O autor destaca como o uso dessas ferramentas estimula o desenvolvimento do pensamento computacional. Este estudo é relevante para esta pesquisa ao demonstrar a importância do design de *prompts* na qualidade das soluções geradas, aspecto que será considerado na comparação entre diferentes modelos de IA.

3.3 Conclusão da Revisão

Esta revisão bibliográfica expõe os principais fundamentos teóricos e estudos recentes sobre o uso da IA na resolução de operações matemáticas complexas com foco nas integrais. As referências analisadas mostram que modelos de linguagem, especialmente quando integrados a motores simbólicos, podem atingir elevados índices de precisão, embora ainda existam limitações significativas. O presente trabalho propõe-se a realizar uma análise comparativa entre diferentes modelos de IA, como GPT-4o, Claude Sonnet 4, Gemini 2.5 Pro, Microsoft Copilot, Grok 3 e DeepSeek R1, considerando critérios como precisão, clareza das explicações, tempo de processamento e outras métricas. Essa proposta busca preencher lacunas identificadas na literatura, oferecendo uma contribuição relevante para o avanço das aplicações de IA no campo da matemática.

4 Metodologia

4.1 Caracterização da Pesquisa

Esta pesquisa caracteriza-se como um estudo comparativo de natureza quantitativa, com abordagem experimental, que visa avaliar a precisão de seis inteligências artificiais generativas distintas na resolução de problemas de cálculo de integrais.

4.2 Seleção das Inteligências Artificiais Generativas

Para este estudo, foram selecionadas seis IAs generativas baseadas nos seguintes critérios:

a) Disponibilidade de acesso: modelos disponíveis publicamente durante o período de coleta de dados (junho de 2025);

b) Capacidades matemáticas declaradas: IAs que expressamente declaram competências em matemática e cálculo;

c) Diversidade arquitetural: representação de diferentes famílias de modelos de linguagem;

d) Atualização recente: modelos com versões lançadas entre 2024 e 2025.

As IAs selecionadas foram:

- GPT-4o (OpenAI): modelo multimodal otimizado para raciocínio matemático;
- Claude Sonnet 4 (Anthropic): modelo com capacidades declaradas em cálculo e análise;
- Gemini 2.5 Pro (Google): modelo com execução de código integrada;
- Microsoft Copilot: modelo baseado em GPT com integração de ferramentas;
- Grok 3 (xAI): modelo com foco em raciocínio avançado;
- DeepSeek R1 (DeepSeek): modelo especializado em matemática e raciocínio formal.

Todas as IAs foram utilizadas em suas configurações padrão, com ferramentas de raciocínio disponíveis quando aplicável.

4.3 Conjunto de Integrais de Teste

Foi selecionado um conjunto de 21 integrais extraídas do livro *Cálculo vol 1. 9 edição* de James Stewart, estratificadas em três níveis de complexidade crescente, conforme classificação apresentada no Quadro 1.

Tabela 1: Classificação das integrais por nível de complexidade

Nível	**Descrição**	**Quantidade**	**Técnicas envolvidas**
Fácil	Integrais definidas básicas	7	Teorema fundamental do cálculo
Médio	Técnicas intermediárias	7	Integração por partes e substituições trigonométricas
Difícil	Técnicas avançadas	7	Frações parciais e integrais impróprias

Fonte: Stewart (2022), adaptado pelo autor (2025).

4.3.1 Integrais do Nível Fácil - Integrais Definidas (Seção 5.2)

As integrais definidas básicas compreendem:

1. $\int_0^2 3x\,dx$

2. $\int_0^3 (5x + 2)\, dx$

3. $\int_1^5 (3x^2 + 7x)\, dx$

4. $\int_0^1 (x^3 - 3x^2)\, dx$

5. $\int_{-2}^5 (10 - 5x)\, dx$

6. $\int_{-4}^3 |\frac{1}{2}x|\, dx$

7. $\int_{-3}^1 (1 + \sqrt{9 - x^2})\, dx$

4.3.2 Integrais do Nível Médio - Integração por Partes e Substituições Trigonométricas

Integração por Partes (Seção 7.1)

1. $\int w \ln w\, dw$

2. $\int (x^2 + 2x) \cos x\, dx$

3. $\int (\ln x)^2\, dx$

4. $\int (1 + x^2) e^{3x}\, dx$

Substituições Trigonométricas (Seção 7.3)

1. $\int x^3 \sqrt{16 + x^2}\, dx$

2. $\int \frac{\sqrt{x^2-1}}{x^4}\, dx$

3. $\int_2^3 \frac{dx}{(x^2-1)^{3/2}}\, dx$

4.3.3 Integrais do Nível Difícil - Técnicas Avançadas

Integrais Impróprias (Seção 7.8)

1. $\int_0^\infty e^{-2x}\, dx$

2. $\int_{-\infty}^0 \frac{z}{z^4+4}\, dz$

3. $\int_2^{14} \frac{dx}{\sqrt[4]{x+2}}\, dx$

Integração por Frações Parciais (Seção 7.4)

1. $\int \frac{5x^4+7x^2+x+2}{x(x^2+1)^2}\,dx$
2. $\int \frac{x^2-3x+7}{(x^2-4x+6)^2}\,dx$
3. $\int \frac{dx}{x\sqrt{x-1}}\,dx$
4. $\int \frac{dx}{x^2+x\sqrt{x}}\,dx$

4.4 Protocolo de Coleta de Dados

4.4.1 Padronização dos Prompts

Foi estabelecido um prompt padrão para garantir condições idênticas de teste para todas as IAs:

> Resolva a seguinte integral, mostrando todos os passos da resolução: [integral]. Inclua a constante de integração quando aplicável e verifique sua resposta por derivação.

4.4.2 Condições de Teste

Para cada integral, foram definidas as seguintes condições experimentais:

- **Número de tentativas**: 3 execuções independentes por IA por integral;
- **Isolamento de contexto**: criação de uma nova conversa com a IA para cada tentativa, evitando influência de contexto anterior;
- **Configurações**: uso das configurações padrão de cada IA;
- **Registro temporal**: anotação do tempo de resposta quando mensurável;
- **Documentação**: captura integral das respostas para análise posterior.

4.4.3 Critérios de Exclusão

Foram estabelecidos os seguintes critérios para exclusão de respostas:

- Recusa da IA em resolver o problema;

- Respostas incompletas ou interrompidas por limitações técnicas;
- Falhas de conectividade durante a coleta.

4.5 Critérios de Avaliação

4.5.1 Método de Avaliação Binária

Devido às limitações práticas de tempo e recursos, foi adotado um sistema de avaliação binária focado na correção do resultado final:

Resposta Correta (1 ponto): A integral foi resolvida corretamente, considerando:

- Resultado final matematicamente correto;
- Presença da constante de integração quando aplicável;
- Limites de integração calculados corretamente para integrais definidas.

Resposta Incorreta (0 pontos): A integral apresenta erro no resultado final, incluindo:

- Resultado matematicamente incorreto;
- Omissão da constante de integração em integrais indefinidas;
- Erros na aplicação de limites de integração;
- Aplicação incorreta de técnicas de integração.

4.5.2 Cálculo da Taxa de Acerto

Para cada modelo de IA, foi calculada a taxa de acerto por categoria:

$$\textbf{Taxa de Acerto por Categoria} = \frac{\text{Número de integrais corretas}}{\text{Número total de integrais da categoria}} \times 100\%$$

$$\textbf{Taxa de Acerto Geral} = \frac{\text{Número total de integrais corretas}}{21} \times 100\%$$

4.6 Validação das Respostas

4.6.1 Padrão de Referência

As soluções de referência foram estabelecidas através de:

- **Gabarito oficial**: respostas fornecidas no livro *Cálculo* de James Stewart;

- **Validação cruzada**: consulta ao Wolfram Alpha e Symbolab para confirmação;
- **Verificação manual**: revisão por especialista em cálculo para casos ambíguos.

4.6.2 Processo de Verificação

Para cada resposta coletada:

1. **Comparação direta**: verificação do resultado final contra o gabarito de referência;
2. **Verificação por derivação**: confirmação através da derivada da resposta obtida (quando aplicável);
3. **Análise de consistência**: verificação da coerência entre as três tentativas de cada modelo.

4.7 Análise Estatística

4.7.1 Estatística Descritiva

Para cada IA foram calculados:

- Taxa de acerto geral e por nível de complexidade;
- Distribuição de frequências de acertos por categoria;
- Medidas de tendência central (média) e dispersão (desvio padrão, coeficiente de variação).

4.7.2 Análise Comparativa

Foi realizada:

- Ranking dos modelos por desempenho geral;
- Análise de correlação entre complexidade da integral e taxa de acerto;
- Identificação de padrões de erro por categoria de complexidade.

4.8 Considerações Éticas e Limitações

4.8.1 Limitações Metodológicas

São reconhecidas as seguintes limitações:

- **Escopo reduzido de avaliação**: foco apenas na correção do resultado final, sem considerar qualidade das explicações ou processos intermediários;
- **Amostra específica**: concentração em problemas de um único livro-texto, podendo não representar toda a diversidade de abordagens em cálculo integral;
- **Variabilidade temporal**: possível inconsistência nas respostas das IAs entre diferentes tentativas;
- **Influência do treinamento**: possível exposição prévia dos modelos aos problemas do livro de Stewart através de dados de treinamento;
- **Diferenças de interface**: variações nas interfaces que podem afetar a entrada de dados.

4.8.2 Simplificação Metodológica

A adoção de um critério binário de avaliação, embora limitando a profundidade da análise, permitiu:

- Maior objetividade na comparação entre modelos;
- Redução de subjetividade na avaliação;
- Viabilidade prática dentro dos recursos disponíveis;
- Foco nos aspectos mais relevantes para usuários finais (correção do resultado).

4.8.3 Reprodutibilidade

Para garantir a reprodutibilidade:

- Todas as integrais e prompts utilizados são apresentados integralmente no documento;
- As versões específicas de cada IA foram documentadas;

- O protocolo detalhado permite replicação por outros pesquisadores;
- Os critérios de avaliação são claramente definidos e objetivos.

4.8.4 Considerações sobre Propriedade Intelectual

Este estudo utiliza apenas funcionalidades públicas das IAs, respeitando os termos de uso de cada plataforma. As respostas geradas são analisadas para fins acadêmicos, enquadrando-se no uso educacional e científico.

5 Cronograma

Esta seção apresenta o cronograma de atividades previstas para o desenvolvimento desta pesquisa, distribuído ao longo de oito semanas. O planejamento temporal foi estruturado de forma a garantir a execução ordenada e eficiente de todas as etapas metodológicas propostas, desde a fase inicial de fundamentação teórica até a apresentação dos resultados finais e conclusões do estudo.

Tabela 2: Cronograma de atividades do projeto

Atividade	**Sem 1-2**	**Sem 3-4**	**Sem 5-6**	**Sem 7-8**
Escolha do tema e desenvolvimento das ideias	✓			
Busca por referências científicas	✓	✓		
Seleção e validação das 21 integrais e desenvolvimento do prompt padronizado		✓		
Apresentação do contexto e objetivos				
Iniciar projeto no LaTeX - Introdução e Resumo			✓	
Redigir a revisão bibliográfica e a metodologia			✓	
Apresentação da Revisão Bibliográfica e Metodologia				▽
Execução dos testes nas IAs				▽
Análise dos resultados e criação dos gráficos				▽
Redigir sobre os resultados e conclusões				▽

Legenda: ✓ = atividade concluída; ▽ = atividade em andamento.

6 Resultados

Esta seção apresenta os resultados obtidos através da avaliação comparativa de cinco sistemas de inteligência artificial generativa na resolução de 21 integrais categorizadas em três níveis de complexidade. Os dados foram coletados seguindo o protocolo metodológico estabelecido, com três execuções independentes para cada integral por sistema de IA.

6.1 Análise Quantitativa Geral

Os resultados demonstram variações significativas no desempenho dos modelos de IA conforme o aumento da complexidade das integrais. A Tabela 3 apresenta um resumo quantitativo consolidado dos resultados obtidos por cada modelo nas três categorias de dificuldade.

Tabela 3: Resumo quantitativo dos resultados por modelo de IA

Modelo de IA	**Integrais Fáceis (%)**	**Integrais Médias (%)**	**Integrais Difíceis (%)**	**Média Geral (%)**
GPT-4o	86	100	43	76.3
Gemini 2.5 Pro	86	100	71	85.7
Claude Sonnet 4	86	100	43	76.3
MS Copilot	86	86	28.5	66.8
Grok 3	86	86	28.5	66.8
DeepSeek R1	86	86	71	81.0

Fonte: Dados da pesquisa (2025).

6.2 Desempenho por Categoria de Complexidade

6.2.1 Integrais Fáceis (Nível 1)

Todos os modelos apresentaram desempenho uniforme nas integrais definidas básicas, obtendo taxa de acerto de 86% (6 de 7 integrais corretas). A Tabela 4 detalha os resultados específicos para esta categoria.

O desempenho uniforme nesta categoria indica que todos os modelos possuem implementação adequada dos algoritmos fundamentais de cálculo integral definido. A única integral que apresentou dificuldades para todos os modelos foi a integral 7, envolvendo uma função com raiz quadrada e valor absoluto, sugerindo limitações específicas no tratamento de funções definidas por partes.

Tabela 4: Desempenho por modelo - Integrais Fáceis

Modelo de IA	**% de Acerto**	**Corretas**	**Principais Limitações**
GPT-4o	86	6/7	Funções definidas por partes
Gemini 2.5 Pro	86	6/7	Funções definidas por partes
Claude Sonnet 4	86	6/7	Funções definidas por partes
Microsoft Copilot	86	6/7	Funções definidas por partes
Grok 3	86	6/7	Funções definidas por partes
DeepSeek R1	86	6/7	Funções definidas por partes

Fonte: Dados da pesquisa (2025).

6.2.2 Integrais Médias (Nível 2)

Na categoria de integrais médias, emergiram dois grupos distintos de desempenho. A Tabela 5 apresenta os resultados detalhados para esta categoria.

Tabela 5: Desempenho por modelo - Integrais Médias

Modelo de IA	**% de Acerto**	**Corretas**	**Principais Limitações**
GPT-4o	100	7/7	Nenhuma identificada
Gemini 2.5 Pro	100	7/7	Nenhuma identificada
Claude Sonnet 4	100	7/7	Nenhuma identificada
Microsoft Copilot	86	6/7	Substituições trigonométricas
Grok 3	86	6/7	Substituições trigonométricas
DeepSeek R1	86	6/7	Substituições trigonométricas

Fonte: Dados da pesquisa (2025).

Os modelos premium (GPT-4o, Gemini 2.5 Pro e Claude Sonnet 4) demonstraram domínio completo das técnicas de integração por partes e substituições trigonométricas, alcançando 100% de acerto. Os demais modelos apresentaram dificuldades específicas em uma integral envolvendo substituição trigonométrica complexa.

6.2.3 Integrais Difíceis (Nível 3)

A categoria de integrais difíceis revelou as maiores disparidades entre os modelos, conforme apresentado na Tabela 6.

Tabela 6: Desempenho por modelo - Integrais Difíceis

Modelo de IA	**% de Acerto**	**Corretas**	**Principais Limitações**
Gemini 2.5 Pro	71	5/7	Frações parciais complexas
DeepSeek R1	71	5/7	Integrais impróprias
GPT-4o	43	3/7	Frações parciais e impróprias
Claude Sonnet 4	43	3/7	Frações parciais e impróprias
Microsoft Copilot	28.5	2/7	Múltiplas técnicas avançadas
Grok 3	28.5	2/7	Múltiplas técnicas avançadas

Fonte: Dados da pesquisa (2025).

Os resultados mostram que Gemini 2.5 Pro e DeepSeek R1 obtiveram desempenho superior, resolvendo corretamente 5 das 7 integrais difíceis. As principais dificuldades identificadas foram relacionadas a integrais envolvendo frações parciais com denominadores complexos e integrais impróprias com limites infinitos.

6.3 Análise Estatística

6.3.1 Medidas de Tendência Central e Dispersão

A Tabela 7 apresenta as estatísticas descritivas por categoria de complexidade.

Tabela 7: Estatísticas descritivas por categoria

Categoria	**Média (%)**	**Desvio Padrão (%)**	**Coef. Variação**	**Amplitude (%)**
Fáceis	86.0	0.0	0.0	0.0
Médias	93.0	8.2	0.09	14.0
Difíceis	47.0	20.1	0.43	42.5

Fonte: Dados da pesquisa (2025).

O coeficiente de variação aumenta progressivamente com a complexidade das integrais, indicando maior diferenciação entre os modelos em problemas avançados. Enquanto as

integrais fáceis apresentaram desempenho uniforme (coeficiente de variação = 0), as integrais difíceis revelaram alta variabilidade (coeficiente de variação = 0.43).

6.3.2 Ranking de Desempenho Geral

Com base na média ponderada dos resultados, estabeleceu-se o ranking apresentado na Tabela 8.

Tabela 8: Ranking geral de desempenho

Posição	Modelo de IA	Média (%)	Principais Forças
1º	Gemini 2.5 Pro	85.7	Excelente em todas as categorias analisadas
2º	DeepSeek R1	81.0	Superior em integrais avançadas testadas
3º	GPT-4o	76.3	Domínio em técnicas intermediárias
3º	Claude Sonnet 4	76.3	Domínio em técnicas intermediárias
5º	Microsoft Copilot	66.8	Competente em problemas básicos
5º	Grok 3	66.8	Competente em problemas básicos

Fonte: Dados da pesquisa (2025).

6.4 Análise Qualitativa dos Erros

Além da análise quantitativa, foram identificados padrões específicos de erros cometidos pelos sistemas de IA:

Erros Procedimentais Comuns:

- Aplicação incorreta de técnicas de integração por partes em funções com múltiplas variáveis
- Dificuldades na decomposição de frações parciais com denominadores de alto grau
- Erros na avaliação de limites em integrais impróprias
- Problemas no tratamento de descontinuidades em funções definidas por partes

Limitações Específicas por Modelo:

- **GPT-4o e Claude Sonnet 4**: Limitações similares em integrais envolvendo radicais complexos
- **Microsoft Copilot e Grok 3**: Dificuldades consistentes em substituições trigonométricas avançadas
- **DeepSeek R1**: Excelente em técnicas algorítmicas, mas ocasionais erros em verificação de resultados
- **Gemini 2.5 Pro**: Melhor desempenho geral, com poucas limitações identificadas

Os resultados indicam que, embora todos os modelos demonstrem competência em cálculo integral básico, existe uma diferenciação significativa nas capacidades de resolução de problemas complexos, sugerindo diferentes níveis de especialização e treinamento em matemática avançada.

7 Conclusões

Este estudo avaliou comparativamente a precisão de cinco sistemas de inteligência artificial generativa na resolução de equações de cálculo integral, utilizando 21 integrais categorizadas por níveis crescentes de complexidade. Os resultados obtidos fornecem insights significativos sobre as capacidades atuais e limitações dos modelos de IA em operações matemáticas avançadas.

7.1 Síntese dos Principais Resultados

A análise dos dados revelou três padrões distintos de desempenho que correlacionam diretamente com a complexidade das operações matemáticas. Em integrais básicas (nível fácil), todos os modelos demonstraram competência uniforme com 86% de acerto, indicando que os algoritmos fundamentais de cálculo integral estão adequadamente implementados em todas as plataformas avaliadas. Esta uniformidade sugere que as técnicas básicas de integração, como o Teorema Fundamental do Cálculo aplicado a funções polinomiais simples, são dominadas universalmente pelos sistemas contemporâneos de IA.

Na categoria intermediária, emergiram dois grupos distintos de desempenho. Os modelos premium (GPT-4o, Gemini 2.5 Pro e Claude Sonnet 4) alcançaram desempenho

perfeito (100% de acerto), demonstrando domínio completo de técnicas como integração por partes e substituições trigonométricas. Em contraste, os modelos alternativos (Microsoft Copilot, Grok 3 e DeepSeek R1) obtiveram 86% de acerto, revelando limitações específicas em substituições trigonométricas complexas.

As integrais avançadas revelaram as maiores disparidades entre os modelos, com variação de desempenho entre 28,5% e 71%. Esta categoria evidenciou que apenas Gemini 2.5 Pro e DeepSeek R1 possuem capacidades robustas para resolver problemas envolvendo frações parciais complexas e integrais impróprias, enquanto os demais modelos apresentaram limitações significativas.

7.2 Recomendações por Aplicação

7.2.1 Modelos Altamente Recomendados

Gemini 2.5 Pro emerge como a escolha superior para aplicações que demandam resolução de integrais complexas. Com média geral de 85,7% e desempenho consistentemente superior em todas as categorias, este modelo demonstra a implementação mais robusta de algoritmos de cálculo integral. É particularmente recomendado para:

- Pesquisa matemática avançada
- Ensino de cálculo em nível superior
- Aplicações industriais que envolvem modelagem matemática complexa
- Verificação de cálculos em engenharia e física teórica

DeepSeek R1 constitui uma alternativa eficaz, especialmente para integrais avançadas, onde obteve desempenho equivalente ao Gemini (71% de acerto). Sua média geral de 81,0% o posiciona como segunda opção recomendada, sendo adequado para:

- Aplicações acadêmicas em matemática aplicada
- Suporte em disciplinas de cálculo avançado
- Desenvolvimento de software matemático

7.2.2 Modelos Recomendados com Restrições

GPT-4o e Claude Sonnet 4 apresentaram desempenho idêntico (76,3% de média geral) e são recomendados para aplicações de complexidade intermediária. Ambos demonstraram excelência em técnicas de integração por partes, mas limitações em problemas avançados. São adequados para:

- Ensino de cálculo básico e intermediário
- Tutoria acadêmica em matemática
- Aplicações educacionais em nível de graduação inicial
- Verificação de cálculos rotineiros em engenharia

7.2.3 Modelos com Limitações Significativas

Microsoft Copilot e Grok 3 obtiveram desempenho inferior (66,8% de média geral) e apresentam limitações substanciais para operações matemáticas avançadas. Embora sejam competentes em problemas básicos, não são recomendados para:

- Pesquisa matemática
- Ensino de cálculo avançado
- Aplicações que exigem alta precisão matemática
- Problemas envolvendo técnicas de integração complexas

Estes modelos podem ser utilizados apenas para verificação de cálculos básicos ou como ferramentas auxiliares em contextos educacionais elementares.

7.3 Padrão de Degradação do Desempenho

A análise estatística revelou um padrão sistemático de degradação do desempenho conforme o aumento da complexidade das integrais. O coeficiente de variação evolui de 0,0 (integrais fáceis) para 0,43 (integrais difíceis), indicando que a complexidade matemática atua como fator diferenciador crítico entre os modelos.

Este fenômeno pode ser explicado por três fatores principais:

Limitações Arquiteturais: Os modelos de linguagem baseados em transformadores, embora eficazes para reconhecimento de padrões, apresentam dificuldades crescentes em operações que exigem raciocínio simbólico rigoroso e manipulação algébrica complexa.

Qualidade dos Dados de Treinamento: A disponibilidade de exemplos de alta qualidade para técnicas avançadas de integração pode ser limitada nos conjuntos de dados de treinamento, resultando em menor precisão para problemas complexos.

Profundidade de Especialização: Apenas os modelos mais avançados incorporam algoritmos especializados para matemática simbólica, explicando a superioridade de Gemini 2.5 Pro e DeepSeek R1 em integrais avançadas.

7.4 Implicações para o Desenvolvimento de IA Matemática

Os resultados sugerem que o desenvolvimento futuro de sistemas de IA para matemática deve focar em três direções principais:

Integração Híbrida: A combinação de modelos de linguagem com sistemas de álgebra computacional (CAS) pode superar as limitações atuais, aproveitando a capacidade explicativa dos LLMs com a precisão dos motores simbólicos.

Especialização Matemática: O treinamento específico em técnicas avançadas de integração e manipulação simbólica pode melhorar significativamente o desempenho em problemas complexos.

Verificação Automática: Implementação de mecanismos de auto-verificação através de derivação ou métodos numéricos pode reduzir erros procedimentais e aumentar a confiabilidade dos resultados.

7.5 Limitações do Estudo e Perspectivas Futuras

Este estudo apresenta limitações que devem ser consideradas na interpretação dos resultados. A amostra de 21 integrais, embora representativa das técnicas fundamentais de cálculo integral, pode não capturar toda a diversidade de problemas matemáticos encontrados na prática. Adicionalmente, a avaliação focou exclusivamente na correção do resultado final, sem considerar a qualidade pedagógica das explicações fornecidas pelos sistemas.

Pesquisas futuras devem expandir o escopo para incluir outras áreas da matemática, como equações diferenciais e cálculo multivariável, além de avaliar aspectos qualitativos

como clareza das explicações e capacidade de detecção de erros. A avaliação longitudinal também é necessária para acompanhar a evolução das capacidades dos modelos ao longo do tempo.

7.6 Considerações Finais

Esta pesquisa demonstra que as inteligências artificiais generativas apresentam capacidades diferenciadas para resolução de integrais, com desempenho que varia significativamente conforme a complexidade dos problemas. Enquanto todos os modelos são adequados para operações básicas, apenas Gemini 2.5 Pro e DeepSeek R1 demonstram competência para problemas avançados.

Os resultados indicam que a escolha apropriada do modelo de IA pode impactar substancialmente a qualidade dos resultados em aplicações matemáticas. Para educadores, pesquisadores e profissionais que utilizam essas ferramentas, é essencial compreender as limitações específicas de cada sistema e aplicá-los em contextos apropriados às suas capacidades.

O campo da IA matemática encontra-se em rápida evolução, e espera-se que desenvolvimentos futuros superem muitas das limitações identificadas neste estudo. No entanto, os padrões observados fornecem uma base sólida para orientar a seleção e aplicação atual dessas tecnologias em contextos educacionais e profissionais que demandam resolução de problemas de cálculo integral.

Referências

[1] LIMA, A. P.; CORSO, C. B. Aplicação de Inteligência Artificial e Modelos Matemáticos para Previsão de Demanda em uma indústria do ramo plástico. *Scientia cum Industria*, v. 8, n. 1, p. 1-12, 2020. Disponível em: https://sou.ucs.br/etc/revistas/index.php/scientiacumindustria/article/view/8122. Acesso em: 5 jun. 2025.

[2] STEWART, James. *Cálculo: Volume 1*. 9. ed. São Paulo: Cengage Learning, 2022.

[3] ZHONG, Yichao et al. MathVista: Evaluating Mathematical Reasoning of Foundation Models. *arXiv preprint*, arXiv:2302.06871, 2023. Disponível em: https://arxiv.org/abs/2302.06871. Acesso em: 5 jun. 2025.

[4] SERVICK, Kelly. Brutal math test stumps AI—not human experts. *Science*, 2024. Disponível em: https://www.science.org/content/article/brutal-math-test-stumps-ai-not-human-experts. Acesso em: 5 jun. 2025.

[5] PEREIRA, Guilherme et al. Updating Calculus Teaching with AI: A Classroom Experience. *Education Sciences*, v. 14, n. 9, p. 1019, 2024. Disponível em: https://www.mdpi.com/2227-7102/14/9/1019. Acesso em: 5 jun. 2025.

[6] SOUZA, Francisco Wallisson Moreira de et al. O ensino e aprendizado matemático com inteligência artificial: uma análise de algumas ferramentas e tecnologias disponíveis. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, v. 10, n. 7, p. 1908-1923, 2024. Disponível em: https://periodicorease.pro.br/rease/article/view/14913. Acesso em: 5 jun. 2025.

[7] AZEVEDO, Greiton Toledo de. Resolução de problemas envolvendo escalas lineares, superficiais e volumétricas: inteligência artificial e pensamento computacional. *Jornal Internacional de Estudos em Educação Matemática*, v. 17, n. 2, p. 138-149, 2024. Disponível em: https://jieem.pgsscogna.com.br/jieem/article/view/12871. Acesso em: 5 jun. 2025.

[8] TRINH, Trieu H. et al. Solving olympiad geometry without human demonstrations. *Nature*, v. 625, p. 476-482, 2024. Disponível em: https://www.nature.com/articles/s41586-023-06747-5. Acesso em: 5 jun. 2025.

[9] DRORI, Iddo et al. Solving Linear Algebra by Program Synthesis. *arXiv preprint*, arXiv:2111.13509, 2021. Disponível em: https://arxiv.org/abs/2111.13509. Acesso em: 5 jun. 2025.

[10] WANG, Jin; FAN, Wenxiang. The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, v. 12, n. 621, 2025. Disponível em: https://doi.org/10.1057/s41599-025-04787-y. Acesso em: 5 jun. 2025.