LLMs e o Raciocínio Lógico: Um Estudo de Caso com Desafios de Lógica

Benjamin Grando Moreira
benjamin.grando@ufsc.br
Universidade Federal de Santa Catarina
Joinville, Santa Catarina, Brasil

ABSTRACT

This paper evaluates the ability of seven Large Language Models (LLMs) to solve logic challenges. The models tested include GPT-4, Claude 3.5 (Sonnet and Haiku), Gemini 1.5, Llama 3.1, Grok, and Mistral 7B. Four challenges were proposed, and the results demonstrate that none of the LLMs were able to solve all challenges correctly. The study highlights the current limitations of LLMs in logical reasoning tasks, despite advancements in other areas of natural language processing.

KEYWORDS

Large Language Models (LLMs), Problem Solving, Artificial Intelligence

1 INTRODUÇÃO

Nos últimos anos, os Modelos de Linguagem de Grande Escala (Large Language Models - LLMs) têm emergido como uma das tecnologias mais promissoras no campo da inteligência artificial, destacando-se por suas capacidades em processamento de linguagem natural e geração de texto. Esses modelos, como o GPT, BERT e outros, têm demonstrado um desempenho impressionante em uma variedade de tarefas, desde tradução automática até criação de conteúdo e assistência em programação. A capacidade desses modelos de compreender e gerar texto de maneira coerente e contextualizada representa um avanço significativo, tornando-os ferramentas valiosas em diversos setores [1, 2].

Contudo, à medida que a aplicação de LLMs se expande, surge a necessidade de avaliar criticamente seu desempenho em tarefas que exigem habilidades de raciocínio lógico. Tais tarefas são fundamentais, pois representam desafios que vão além da simples associação de palavras, exigindo uma compreensão mais profunda e a habilidade de conectar conceitos de maneira lógica. A avaliação comparativa do desempenho dos LLMs em desafios de lógica é crucial para identificar suas limitações e potencialidades, orientando o desenvolvimento de modelos futuros que possam superar essas barreiras.

Neste contexto, o presente estudo se propõe a comparar o desempenho de LLMs distintos em quatro desafios de lógica, com o objetivo de identificar quão bem cada modelo se sai na resolução dessas tarefas. A importância desta análise reside na capacidade de discernir as forças e fraquezas de cada modelo, fornecendo insights valiosos para pesquisadores e desenvolvedores interessados em aprimorar as capacidades de raciocínio lógico dos LLMs.

A literatura atual sugere que, apesar dos avanços, muitos modelos ainda enfrentam dificuldades significativas em tarefas que exigem raciocínio lógico complexo [3–5]. A resolução de desafios de lógica representa uma das áreas mais complexas e intrigantes dentro do campo da inteligência artificial. Tais desafios exigem que os modelos não apenas compreendam a linguagem em um nível superficial, mas também que realizem inferências, estabeleçam conexões entre conceitos e apliquem regras lógicas de maneira precisa. Essa complexidade inerente aos desafios de lógica se traduz em um obstáculo significativo para os LLMs, que, apesar de seus avanços notáveis, ainda enfrentam limitações substanciais quando confrontados com problemas que demandam raciocínio lógico sofisticado.

Os resultados preliminares deste estudo indicam que nenhuma das LLMs avaliadas conseguiu resolver corretamente todos os quatro desafios de lógica, destacando a necessidade contínua de pesquisa e inovação nesta área. Ao identificar as lacunas no desempenho dos modelos atuais, este trabalho contribui para o avanço do campo, incentivando o desenvolvimento de LLMs mais robustos e capazes de lidar com a complexidade do raciocínio lógico.

Um total de 7 modelos de LLMs foram avaliados neste trabalho, cada um buscando resolver 4 desafios lógicos distintos, sendo os modelos e desafios descritos na Seção 3.

2 JUSTIFICATIVA

Embora os LLMs tenham demonstrado capacidades impressionantes em tarefas que envolvem processamento de linguagem natural, como tradução e geração de texto, seu desempenho em tarefas que requerem lógica e raciocínio crítico ainda está aquém das expectativas. Estudos como [3, 4] indicam que esses modelos, embora eficientes em capturar padrões estatísticos da linguagem, frequentemente falham em aplicar lógica de maneira consistente. Essa limitação é particularmente evidente em cenários onde a solução de problemas depende da capacidade de entender premissas complexas e derivar conclusões válidas a partir delas.

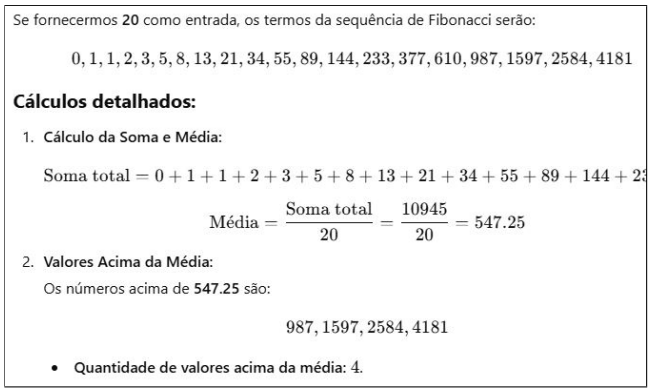
Mesmo em questões mais simples, uma LLM pode apresentar resoluções incorretas a problemas. Na Figura 1 é mostrado um trecho de conversa com o ChatGPT-4o em uma questão relacionada com a sequência de Fibonacci, solicitando quantos valores estão acima da média dos primeiros 20 termos da sequência. A resposta indicada é 5, embora tenha sido retornada a resposta como 4.

É importante trazer que o mesmo LLM, quando questionado na mesma interação "Sejam os valores 0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, 4181. Quantos desses valores estão acima da média desses valores?", é obtido o resultado correto.

3 MATERIAIS E MÉTODOS

Neste trabalho, sete LLMs foram avaliados, 5 deles comercais e 2 de código aberto. O modelos foram os seguintes:

- GPT-4o: modelo da empresa OpenAI;
- Claude 3.5 Sonnet: modelo da empresa Anthropic;
- Claude 3.5 Haiku: versão mais rápida da Claude 3.5 Sonnet;


Se fornecermos **20** como entrada, os termos da sequência de Fibonacci serão:

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, 4181$$

Cálculos detalhados:

1. **Cálculo da Soma e Média:**

   $\text{Soma total} = 0 + 1 + 1 + 2 + 3 + 5 + 8 + 13 + 21 + 34 + 55 + 89 + 144 + 2$

   $$\text{Média} = \frac{\text{Soma total}}{20} = \frac{10945}{20} = 547.25$$

2. **Valores Acima da Média:**

   Os números acima de **547.25** são:

   $$987, 1597, 2584, 4181$$

   - **Quantidade de valores acima da média:** 4.

Figura 1: Erro de resolução apresentado pelo GPT-4o

- Gemini 1.5: modelo do Google;
- Llama 3.1: modelo da Meta e de código aberto;
- Grok: modelo da empresa xAI; e
- Mistral 7B: modelo de código aberto.

Os modelos GPT-4o, Claude 3.5 Sonnet, Claude 3.5 Haiku, Gemini 1.5, Llama 3.1, Grok e Mistral 7B serão referidos no decorrer deste trabalho como, respectivamente, GPT4, ClaudeS, ClaudeH, Gemini, Llama, Grok e Mistral.

Foram elaborados 4 problemas que envolvem raciocínio lógico para serem avaliados por cada uma das LLMs. A resposta esperada para cada uma das questões é apresenta na Seção 5. Os desafios lógicos são:

(1) Um elefante incomoda muita gente, dois elefantes incomodam incomodam muito mais. Três elefantes incomodam muita gente, quadro elefantes incomodam incomodam incomodam incomodam muito mais. Cada seria a continuação do texto para cinco e seis elefantes?
(2) Imagine uma codificação simples, de forma que a palavra "Artigo"seja codificada como "Bsujhp". Como ficaria codificada a palavra "Paper"?
(3) Se janeiro é 17, fevereiro é 49, março é 95, então maio é?
(4) Indique qual alternativa representa a solução da operação 3 + 3 x 5: a) 16; b) 20; c) 30; d) 45.

Os quarto item não é realmente uma questão de lógica e sim de cálculo (mais especificamente uma questão de prioridade dos operadores), mas como não é fornecida uma opção válida, espera-se avaliar essa discordância.

Para início das interações, o seguinte prompt foi utilizado: “resolva alguns desafios de lógica a seguir e apresente como você chegou a conclusão da solução”. Esse prompt é utilizado para obter tanto a resolução para o desafio, quanto uma explicação sobre o processo de avaliação.

Para cada erro, foi revista a questão fornecendo alguma complementação (com exceção para o quarto desafio). O objetivo foi tentar ajudar o modelo na elaboração da solução. As seguintes complementações aos desafios foram realizadas:

(1) Adicionado “Cinco elefantes incomodam muita gente, seis elefantes incomodam incomodam incomodam incomodam incomodam incomodam incomodam muito mais ao trecho e solicitado a continuação para sete e oito elefantes.
(2) Foi alterado para utilizar a palavra *paper* utilizando somente letras minúsculas. Também foi adicionada a codificação de Paper como Qbqfs, sendo solicitada a codificação da palavra *Revista*.
(3) Adicionado que “abril é 165”.

É importante destacar que em nenhuma situação incorreta foi informado que a resposta estava errada, apenas solicitado um novo desafio com a complementação.

4 RESULTADOS

Para apresentação dos resultados, inicialmente é mostrado uma visão geral sobre os acertos e erros de cada LLM em cada uma dos desafios. Em seguida, uma explicação apresenta os erros cometidos em cada avaliação incorreta das LLMs, bem como o resultado da aplicação da complementação em cada desafio incorreto. A Tabela 1 sumariza os resultados considerando a aplicação somente dos desafios originais (sem as complementações).

LLM	1	2	3	4
GPT4	👍	👍	👎	👍
ClaudeH	👎	👍	👎	👍
ClaudeS	👍	👍	👎	👍
Gemini	👎	👎	👎	👎
Llama	👍	👍	👎	👍
Grok	👍	👍	👎	👍
Mistral	👍	👍	👎	👍

Tabela 1: Resultado da avaliação das LLMs para os 4 desafios propostos

Fica evidenciado na Tabela 1 que todas os LLMs erram em responder ao terceiro desafio. Boa parte dos LLMs acertou os demais desafios, embora o Gemini não tenha conseguido acertar nenhum deles.

O GPT4 errou sua resposta ao terceiro desafio, mas indicou que “[...]Ainda há uma dificuldade em encontrar um padrão direto.[...]Para descobrir a solução correta, detalhes adicionais ou dicas seriam necessários ou uma revisão das possíveis alternativas numéricas”. Quando fornecida a complementação para a questão, mesmo assim não foi possível resolver.

O ClaudeH errou o primeiro desafio indicando como trecho, incorretamente, “Cinco elefantes incomodam incomodam incomodam incomodam incomodam muito mais”. Quando o texto foi ampliado com a complementação, o ClaudeH conseguiu acertar o primeiro desafio. No erro do terceiro desafio, embora uma resposta diferente da do autor tenha sido apresentada, foi apresentada uma proposta de solução válida.

Quanto ao ClaudeS errou de maneira similar ao ClaudeH, com a diferença que no primeiro desafio, embora não tenha respondido realmente a questão, é apresentada uma avaliação correta para o desafio, inclusive associando o texto a cantiga infantil que inspirou o desafio.

O Gemini, que errou todos os desafios, mesmo com a complementação continuou não apresentando uma resposta correta. Destaca-se que, no segundo desafio, um erro na avaliação da letra *p* minúscula foi identificado (a letra foi trocada para *i*), mesmo ele explicando que "[...]cada letra em "Artigo"é substituída pela letra seguinte no alfabeto[...]". No quarto desafio, o Gemini inclusive gerou a afirmação de que "[...]3 + 15 = 16".

Em relação ao Llama, esse somente não acertaram o terceiro desafio, sendo o erro de avaliação semilar aos demais.

Quanto ao Grok, embora no quarto desafio a ferramenta tenha indicado que a alternativa *A* era a correta, em toda a explicação ele afirmou que o resultado era 18. Para O terceiro desafio, seu erro foi similar aos demais.

Por fim, o Mistral, embora tenha errado no terceiro desafio, esse foi o se aproximou no quesito de considerar que a parte inicial do valor vem da posição do mês elevado ao quadrado, embora não tenha identificado a parte final.

5 RESPOSTAS ESPERADAS

O primeiro desafio é inspirado em uma cantiga infantil (ou pode ser considerado um trecho dessa cantiga). O resultado esperado é que "Cinco elefantes incomodam muita gente, seis elefantes incomodam incomodam incomodam incomodam incomodam incomodam incomodam muito mais". Ou seja, em número ímpares aparece uma ocorrência da palavra *incomodam*, enquanto em números pares a palavra *incomodam* aparece em quantidade igual ao do número. O desafio não é apenas uma correspondência simples entre o número e a quantidade de repetições, tornando mais desafiador identificar o padrão.

Para o segundo desafio, a codificação da palavra *Paper* resulta em *Qbqfs*, sendo que cada letra é substituída pela sua próxima letra do alfabeto.

O terceiro desafio era o mais complexo. A resposta para o mês de maio é 254, sendo o valor composto da posição do mês ao quadro e do número de letras que compõem o nome do mês. No caso, maio é o quinto mês ($5^2 = 25$) e possui 4 letras, resultando na concatenação de 25 e 4, ou seja, 254. As LLMs buscaram identificar alguma sequência numérica diretamente associada com elementos dos meses, seja sua quantidade de letras, posição no mês, atribuição de valores para as letras, mas um número gerado por um casting da concatenação de dois valores numéricos foi desafiador demais para as LLMs, mesmo cada um dos cálculos sendo simples.

No quarto desafio, o resultado do cálculo é 18, embora um erro relacionado à prioridade dos operadores poderia conduzir a resposta 45. De qualquer maneira, a resposta correta não é uma das opções válidas, sendo necessário indicar que não existe uma alternativa com a resposta correta.

6 CONSIDERAÇÕES FINAIS

Este estudo comparativo do desempenho de sete LLMs em desafios de lógica revelou limitações na capacidade desses modelos de raciocinar logicamente. Apesar dos avanços notáveis em áreas como geração de texto e tradução, a aplicação consistente de regras lógicas e a capacidade de inferência em contextos complexos ainda representam um desafio para os LLMs atuais. Nenhuma das LLMs avaliadas, incluindo modelos comerciais e de código aberto, conseguiu resolver todos os quatro desafios propostos, indicando a necessidade de aprimoramentos no desenvolvimento desses modelos.

A dificuldade encontrada pelas LLMs em resolver o terceiro desafio, que envolvia um padrão numérico complexo, sugere que a dependência de padrões estatísticos na linguagem pode ser insuficiente para lidar com problemas que exigem raciocínio lógico mais profundo. Os resultados obtidos reforçam a importância da avaliação contínua e comparativa de LLMs em diferentes tipos de tarefas, especialmente aquelas que demandam habilidades cognitivas mais complexas. A pesquisa futura nessa área deve se concentrar no desenvolvimento de métodos que permitam aos LLMs superar essas limitações, explorando abordagens que vão além da simples aprendizagem estatística de padrões linguísticos. Os resultados também reforçam a necessidade de, em alguns momentos, ser necessário fornecer mais informações para que o LLM consiga alcançar o resultado necessário.

Destaca-se que os resultados revelam os estados atuais experimentados em cada uma das LLMs, sendo possível obter outras respostas no futuro. Por esse motivo, as conversas completas estabelecidas com cada LLM estão disponibilizadas em https://bit.ly/llms-logica, para que seja possível obter mais detalhes das respostas e realizar comparações futuras com as novas versões das LLMs.

REFERÊNCIAS

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[3] Gary Marcus and Ernest Davis. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon Books, USA, 2019. ISBN 1524748250. doi: 10.5555/3364958.

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922.

[5] Saumya Malik. Lost in the logic: An evaluation of large language models' reasoning capabilities on lsat logic games. *ArXiv*, abs/2409.19012, 2024.