Evaluating the Ability of ChatGPT and DeepSeek to Solve Propositional Logic Proofs Using the Analytic Tableau Deductive System

Taís Rodrigues Sandes[1], **Davi Romero de Vasconcelos** [1]
Maria Viviane de Menezes[1], **Victória de Oliveira Lima** [1]

[1]Universidade Federal do Ceará - Campus Quixadá

{tais.sandes,victoriaoliveiral}@alu.ufc.br, {daviromero,vivianemenezes}@ufc.br

Abstract. *Large Language Models (LLMs) have been widely applied in educational contexts but still face challenges in tasks requiring rigorous logical reasoning. This paper evaluates the performance of ChatGPT-4o and DeepSeek R1 (DeepThink) in solving Propositional Logic exercises using the Analytic Tableau deductive system. The responses were analyzed based on the correct application of tableau rules. Results show that although DeepSeek outperformed ChatGPT in the number of correct answers, both models still exhibit significant limitations, especially in proofs involving multiple rule applications.*

Resumo. *Os Modelos de Linguagem de Grande Escala (LLMs) têm sido amplamente aplicados em contextos educacionais, mas enfrentam desafios em tarefas que exigem raciocínio lógico rigoroso. Este artigo avalia o desempenho dos modelos ChatGPT-4o e DeepSeek R1 (DeepThink) na resolução de exercícios de Lógica Proposicional, utilizando o sistema dedutivo Tableau Analítico. As respostas foram analisadas com base na aplicação correta das regras do sistema. Os resultados mostram que, embora o DeepSeek tenha superado o ChatGPT em número de acertos, ambos os modelos ainda demonstram limitações significativas, especialmente em provas com múltiplas aplicações de regra.*

1. Introdução

Os Modelos de Linguagem de Grande Escala (Large Language Models – LLMs) têm ampliado significativamente as aplicações da inteligência artificial, especialmente em tarefas de processamento de linguagem natural (PLN) [Viegas et al. 2024]. Devido à sua capacidade de processar grandes volumes de texto e adaptar-se a diferentes contextos, esses modelos vêm despertando o interesse de pesquisadores e instituições educacionais, sobretudo pelo seu potencial para apoiar o ensino personalizado [Kasneci et al. 2023, Tlili et al. 2023]. No entanto, seu desempenho em domínios que exigem raciocínio lógico estruturado ainda levanta questionamentos [Liu et al. 2023, Martins et al. 2025].

A Lógica para Computação é uma disciplina abordada em grande parte dos cursos de graduação na área de Tecnologia da Informação. Entre os principais conteúdos abordados está a Lógica Proposicional [Huth 2004]. Nesse contexto, destaca-se o sistema dedutivo *Tableau Analítico*, um método baseado na refutação que testa a validade de argumentos assumindo as premissas como verdadeiras e a conclusão como falsa, buscando assim a contradição [Vasconcelos 2023].

Estudos como o de [Saparov et al. 2023] investigaram a capacidade de LLMs (FLAN-T5 [Chung et al. 2024], LLaMA [Touvron et al. 2023], GPT-3.5 [OpenAI 2021] e PaLM [Chowdhery et al. 2023]) em realizar provas de dedução natural. Os resultados indicam que, embora os modelos sejam capazes de aplicar regras simples após receber contexto, enfrentam dificuldades na generalização composicional, em provas complexas.

Neste contexto, este artigo tem como objetivo avaliar a habilidade dos modelos ChatGPT-4o e DeepSeek na resolução de exercícios de Lógica Proposicional utilizando o sistema dedutivo *Tableau Analítico*. Diferentemente de [Saparov et al. 2023], nosso estudo utiliza como entrada fórmulas escritas em LaTeX, extraídas de uma base de dados aplicada em uma disciplina de Lógica para Computação. As respostas foram analisadas quanto à correção lógica e ao uso adequado das regras do sistema. Até onde sabemos, este é o primeiro estudo a investigar LLMs nesse tipo tipo de prova.

O restante desse artigo está organizado da seguinte forma. a Seção 2 apresenta a fundamentação teórica; a Seção 3 discute trabalhos relacionados; a Seção 4 descreve a metodologia adotada; a Seção 5 apresenta os resultados obtidos; e a Seção 6 traz as conclusões e direções para trabalhos futuros.

2. Fundamentação

2.1. Lógica Proposicional

A Lógica Proposicional é uma linguagem formal voltada à representação e análise de proposições e suas relações lógicas. Constituída por um alfabeto, basea-se em proposições ou frases declarativas sobre as quais se pode argumentar a veracidade ou falsidade. As frases declarativas consideradas atômicas, como "*o dólar subiu*" são chamadas de átomos proposicionais. O conjunto desses átomos, juntamente com os conectivos (negação $\neg$, conjunção $\wedge$, disjunção $\vee$ e implicação $\rightarrow$) e os símbolos '(', ')', compõem o alfabeto da Lógica Proposicional. Essa linguagem pode ser definida por meio de uma gramática na forma de *Backus Naur* (BNF - *Backus Naur Form*) onde $\bot$ representa a contradição e P qualquer átomo proposicional. Além disso, cada ocorrência φ a direita de '::=' representa qualquer fórmula já construída. como a seguir [Huth 2004]:

$$\varphi ::= \bot \mid P \mid (\neg\varphi) \mid (\varphi \wedge \varphi) \mid (\varphi \vee \varphi) \mid (\varphi \rightarrow \varphi)$$

Por exemplo, podemos usar o átomo proposicional P para representar a frase "*está chovendo*" e o átomo proposicional Q para representar "*as ruas estão molhadas*". Para expressar frases mais complexas, utilizamos conectivos lógicos. Por exemplo: $\neg P$ codifica a frase "não está chovendo"; $P \wedge Q$ codifica a frase "está chovendo *e* as ruas estão molhadas"; $P \vee Q$ codifica a frase "está chovendo *ou* as ruas estão molhadas"; e $P \rightarrow Q$ representa "*se* está chovendo, "*então* as ruas estão molhadas".

2.2. Tableau Análitico

O *Tableau Analítico* é um método que opera com base na refutação e busca contra-exemplos para estabelecer a validade das fórmulas, sendo aplicável principalmente à lógica proposicional [Vasconcelos 2023]. Para demonstrar que $\Gamma \vdash \varphi$, assumimos inicialmente que todas as fórmulas em Γ são verdadeiras e que φ é falsa. A partir dessa suposição, buscamos encontrar uma contradição, o que provaria que nossa suposição sobre φ ser falsa está incorreta, confirmando assim $\Gamma \vdash \varphi$. Caso não encontremos, isso

indica que em pelo menos uma situação Γ é verdadeiro e φ é falso, produzindo contra-exemplo.

Para determinar a veracidade ou falsidade de uma fórmula, o método utiliza os símbolos T para verdade e F para falsidade. Apartir do tableau inicial, aplicam-se regras de expansão que podem adicionar novas fórmulas ao final de um ramo, conhecidas como regras do tipo α, ou dividem um ramo em dois, conhecidas como β (Figura 1).

Tipo α	$T\ \varphi \wedge \psi$ | $T\ \varphi$ $T\ \psi$	$F\ \varphi \vee \psi$ | $F\ \varphi$ $F\ \psi$	$F\ \varphi \to \psi$ | $T\ \varphi$ $F\ \psi$	$T\ \neg\varphi$ $\quad$ $F\ \neg\varphi$ | $\quad$ | $F\ \varphi$ $\quad$ $T\ \varphi$
Tipo β	$F\ \varphi \wedge \psi$ $F\ \varphi$ $\quad$ $F\ \psi$	$T\ \varphi \vee \psi$ $T\ \varphi$ $\quad$ $T\ \psi$	$T\ \varphi \to \psi$ $F\ \varphi$ $\quad$ $T\ \psi$	

Figura 1. Regras de expansão

Em cada ramo do tableau, uma fórmula só pode ser expandida uma vez, caracterizando a Expansão Única. Um ramo é considerado saturado quando não há mais fórmulas para expandir. Se um ramo contém um par de fórmulas T φ e F φ, ele é fechado e não requer mais expansão. Um tableau é fechado quando todos os seus ramos estão fechados, indicando que $\Gamma \vdash \varphi$. Caso um ramo esteja saturado, mas não fechado, ele serve como um contraexemplo, demonstrando que $\Gamma \vdash \varphi$ não é válido [Vasconcelos 2023]. Na (Figura 2) é mostrado um exemplo de prova utilizando o método de Tableau Análitico.

$$
\begin{array}{c}
T\ A \to B \\
T\ \neg B \\
F\ \neg A \\
| \\
F\ B \\
| \\
T\ A \\
\diagup \quad \diagdown \\
F\ A \quad T\ B \\
| \quad\quad | \\
\times \quad\quad \times
\end{array}
$$

Figura 2. Aplicação do tableau para $A \to B,\ \neg B \vdash \neg A$

Para representar as árvores de derivação construídas pelo método, utilizamos o pacote LaTeX qtree. Esse pacote permite expressar a estrutura hierárquica dos ramos por meio de notação entre colchetes, no comando \Tree, em que cada nó pode conter expressões matemáticas renderizadas com `$...$`, e múltiplas fórmulas podem ser agrupadas em um único nó com a separação de linhas via \\. Nós terminais fechados são indicados por \times. A organização e o espaçamento da árvore são automaticamente ajustados pelo pacote, assegurando clareza na visualização da estrutura dedutiva [Siskind and Dimitriadis 2008].

2.2.1. Regra da Negação

A regra da negação é representada na Figura 3. Quando um nó contém a fórmula $T\ \neg\varphi$, ele se expande para um nó filho com $F\ \varphi$, mantendo-se na mesma ramificação. Já se

o nó inicial contém $F\ \neg\varphi$, ele gera um filho com $T\ \varphi$, igualmente na mesma linha de descendência. Em ambos os casos, a estrutura da árvore não se ramifica, pois a negação envolve apenas uma alternativa sem ambiguidade.

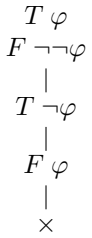

Figura 3. Aplicação da regra Negação.

2.2.2. Regra da Conjunção

A regra da conjunção verdade, ilustrada na Figura 4, aplica-se quando um nó contém $T\ (\varphi \wedge \psi)$. Nesse caso, o nó gera dois filhos sucessivos, $T\ \varphi$ seguido de $T\ \psi$, ambos pertencentes à mesma linha descendente. Isso reflete a exigência semântica de que, para que uma conjunção seja verdadeira, ambas as subfórmulas também devem ser verdadeiras.

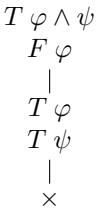

Figura 4. Aplicação da regra Conjunção Verdade.

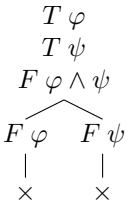

Figura 5. Aplicação da regra Conjunção Falsa.

Por outro lado, a regra da conjunção falsa, mostrada na Figura 5, produz uma bifurcação: a fórmula $F\ (\varphi \wedge \psi)$ gera dois ramos distintos, um contendo $F\ \varphi$ e o outro contendo $F\ \psi$. Essa separação representa a ideia de que, para que uma conjunção seja falsa, basta que ao menos uma das partes seja falsa — mas o tableau deve explorar ambas as possibilidades.

2.2.3. Regra da Disjunção

A disjunção verdade, conforme apresentado na Figura 6, também dá origem a uma bifurcação. A presença de $T\ (\varphi \vee \psi)$ leva à criação de dois ramos: um com $T\ \varphi$ e outro com $T\ \psi$. Essa separação acontece porque, para que uma disjunção seja verdadeira, basta que pelo menos uma das partes também seja. Por isso, o tableau analisa cada possibilidade em um ramo diferente.

Em contraste, a disjunção falsa, ilustrada na Figura 7, não gera ramificação. A fórmula $F\ (\varphi \vee \psi)$ se expande com dois nós consecutivos no mesmo ramo, contendo $F\ \varphi$ e $F\ \psi$. Isso corresponde à necessidade de que ambas as subfórmulas sejam falsas para que a disjunção o seja.

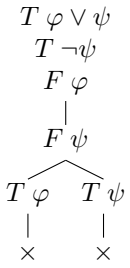

Figura 6. Aplicação da regra Disjunção Verdade.

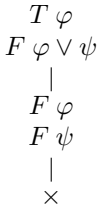

Figura 7. Aplicação da regra Disjunção Falsa.

2.2.4. Regra da Implicação

A implicação verdade, representada na Figura 8, conduz a uma bifurcação: o nó com $T\ (\varphi \to \psi)$ gera dois ramos, um contendo $F\ \varphi$ e o outro contendo $T\ \psi$. Essa ramificação traduz o fato de que a implicação é verdadeira sempre que a hipótese é falsa ou a conclusão é verdadeira — e, portanto, deve-se explorar ambas as possibilidades.

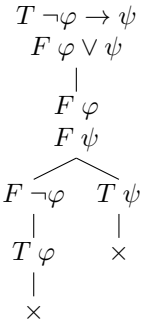

Figura 8. Aplicação da regra Implicação Verdade.

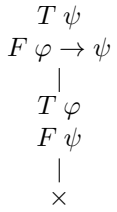

Figura 9. Aplicação da regra Implicação Falsa.

Por fim, a implicação falsa, conforme a Figura 9, expande-se de forma linear: a presença de $F\ (\varphi \to \psi)$ gera dois nós consecutivos no mesmo ramo, $T\ \varphi$ seguido de $F\ \psi$. Essa estrutura corresponde ao único caso em que uma implicação é falsa, quando a hipótese é verdadeira e a conclusão é falsa simultaneamente.

2.3. Large Language Models (LLMS)

Os Modelos de Linguagem de Grande Escala (LLMs) representam um avanço no campo da inteligência artificial (IA) e do processamento de linguagem natural (NLP). São sistemas de IA baseados em aprendizado de máquina [Carbonell et al. 1983] e treinados em grandes volumes de dados para identificar padrões de linguagem e, como consequência, serem capazes de compreender e gerar linguagem natural de forma coerente [Naveed et al. 2023]. A evolução dos LLMs tem crescido nos últimos anos, alguns exemplos dos modelos são do GPT-3 (Generative Pretrained Transformer) [Brown et al. 2020], ELMo(Embeddings from Language Model) [Sarzynska-Wawer et al. 2021], e outros mais recentes como o GPT-4,PALM2 [Anil et al. 2023] e LLaMa [Touvron et al. 2023].

2.3.1. ChatGPT

O ChatGPT, desenvolvido pela OpenAI [OpenAI 2021], tem sido amplamente utilizado para auxiliar no processo de aprendizagem permitindo que educadores identifiquem e compreendam erros recorrentes nesse processo. Baseado em grandes modelos de linguagem, o ChatGPT é impulsionado por versões aprimoradas da arquitetura GPT [Martins et al. 2025]. O GPT-3, lançado em 2020, foi reconhecido pelo MIT Technology Review como uma das " *Top 10 Breakthrough Technologies´´* em 2021, devido à sua capacidade de modelagem avançada, generalização multitarefa e aprendizado eficiente com poucas tentativas [Zhang and Li 2021]. Com a introdução do GPT-3.5 em 2022, foram implementadas melhorias significativas em desempenho e segurança. Em 2023, o GPT-4 trouxe avanços com o uso de modelos de recompensa baseados em regras e aprendizado por reforço com feedback humano, o que aprimorou ainda mais a precisão e a segurança das respostas em relação às versões anteriores [Koubaa 2023].

2.3.2. DeepSeek

O DeepSeek foi criado por uma startup chinesa e surgiu como uma IA essencial para o treinamento de modelos de larga escala e com menor custo e benefício do mercado [Aydin et al. 2025]. Alguns estudos destacam os avanços que tornam o treinamento de modelos de IA mais eficientes, dentro desse contexto, alguns dos modelos mais relevantes incluem DeepSeekMath, DeepSeek-V2, DeepSeek-V3 e DeepSeekR1 [Shao et al. 2024, Liu et al. 2024, Liu et al. 2024, Guo et al. 2025], todos trazendo inovações significativas. O DeepSeek-R1, em particular, representa um avanço no raciocínio automatizado, adotando uma abordagem baseada em regras que melhora o alinhamento do modelo com tarefas de matemática, codificação e raciocínio [Aydin et al. 2025, Guo et al. 2025].

3. Trabalhos Relacionados

O estudo de [Martins et al. 2025] investigou a capacidade do modelo ChatGPT em resolver exercícios de Dedução Natural em Lógica Proposicional e Lógica de Predicados. Foram utilizados 41 exercícios de lógica proposicional e 20 de lógica de predicados, e os resultados indicaram um desempenho insatisfatório do modelo, com apenas 6 acertos (14,63%) em lógica proposicional e 3 (15%) em lógica de predicados. Os erros foram classificados em dois tipos: erros lógicos, relacionados à aplicação incorreta das regras de dedução, e erros de referência, ligados à citação inadequada de linhas. A maioria dos erros foi de natureza lógica, sugerindo que, apesar de suas capacidades linguísticas, o modelo ainda encontra dificuldades em tarefas que exigem raciocínio formal rigoroso.

Já em [Saparov et al. 2023], os autores exploraram a habilidade de generalização dos LLMs em provas de dedução, com foco especial na chamada generalização composicional que é a capacidade de aplicar múltiplas regras em sequência. O estudo mostrou que os modelos podem surpreender positivamente quando expostos a exemplos variados, mas também evidenciou limitações importantes, como a dificuldade de aplicar certas regras sem demonstrações explícitas. Outro ponto relevante foi a observação de que modelos menores, quando bem ajustados, podem ter desempenho comparável ao de modelos maiores, questionando a correlação direta entre tamanho do modelo e eficácia.

O trabalho de [Lalwani et al. 2024] apresenta o sistema NL2FOL, que combina

modelos de linguagem com lógica formal para detectar falácias em argumentos expressos em linguagem natural. A abordagem propõe a tradução de sentenças naturais para fórmulas em Lógica de Primeira Ordem, posteriormente avaliadas por um solver SMT. Embora o modelo tenha apresentado resultados promissores, os autores destacam desafios como a ambiguidade da linguagem natural e a complexidade da conversão lógica em alguns casos. O estudo propõe aprimoramentos futuros, incluindo o uso de modelos mais avançados e a incorporação de contexto semântico na tradução.

Esses trabalhos estão relacionados ao presente estudo por investigarem a aplicação de LLMs em tarefas de raciocínio lógico. No entanto, diferem quanto ao enfoque metodológico, enquanto os estudos anteriores abordam dedução natural ou tradução para lógica formal, este trabalho utiliza o método Tableau Analítico para avaliar a capacidade de LLMs na resolução de exercícios de lógica proposicional em um contexto educacional.

4. Metodologia

Para avaliar o desempenho dos modelos na construção de provas lógicas com o método Tableau Analítico, foi aplicado um conjunto de 40 questões de Lógica Proposicional, utilizadas em disciplinas de Lógica para Computação em cursos de graduação em Tecnologia da Informação. Esse conjunto reflete desafios práticos típicos enfrentados por alunos. O experimento foi conduzido em três etapas: (i) elaboração de uma base de dados com perguntas e respostas formuladas no sistema Tableau Analítico (https://bit.ly/3G7UBVL); (ii) submissão das questões aos modelos com o devido contexto (https://bit.ly/44tvkyv; https://bit.ly/4k4v7GH); e (iii) avaliação das respostas com base em erros lógicos, como falhas na aplicação das regras ou incoerências na estrutura dos argumentos.

Foram selecionados quatro modelos para o experimento: o GPT-4o e a versão reflexiva, e DeepSeek V3 e sua versão reflexiva (DeepSeek R1), A escolha do GPT-4o se justifica por seu amplo uso nos trabalhos relacionados a esse estudo [Martins et al. 2025, Saparov et al. 2023], enquanto o DeepSeek foi incluído por ser uma alternativa gratuita ao ChatGPT, com potencial competitivo, especialmente o modelo R1, pois foi desenvolvido com foco em tarefas complexas de raciocínio, matemática e programação.

Os exercícios foram submetidos como textos em LaTeX seguindo o padrão apresentando por [Vasconcelos 2023] (e.g, `$A \land B, \vdash \lnot (A \rightarrow \lnot B)$`) correspondentes a teoremas com premissas e conclusão:

$A \land B \vdash \neg(A \rightarrow \neg B)$ (com premissa $A \land B$ e conclusão $\neg(A \rightarrow \neg B)$)

A comunicação com os modelos foi realizada por meio da API (Application Programming Interface) disponibilizada pela OpenAI, acessada por intermédio de sua biblioteca oficial para a linguagem Python (https://api.openai.com/v1/chat/completions). As interações seguiram as diretrizes de boas práticas propostas pela própria OpenAI. Foi adotado o modelo mais recente disponível, conforme a recomendação de priorizar versões atualizadas. As instruções foram posicionadas no início do prompt e devidamente separadas do contexto por delimitadores como as aspas triplas, conforme sugerido. Buscou-se ainda fornecer descrições claras, específicas e detalhadas acerca do objetivo da tarefa, do formato esperado das respostas e do estilo desejado. Além disso, foram apresentados exemplos de saída para orientar o modelo quanto ao formato de resposta esperado. Quanto à abordagem de engenharia de prompt, embora a técnica fine-tuning não tenha

sido utilizada neste experimento, há a intenção de emprega-la em trabalhos futuros, a fim de aprofundar a personalização do comportamento do modelo.

Na (Figura 10) e (Figura 11) são mostrados respectivamente o enunciado e a resposta exemplo fornecido ao modelo. Em seguida eram aplicadas as perguntas, que foram inseridas conforme representado na (Figura 12).

```
“escreva a prova A \lor (B \land C) \vdash (A \lor B)\land (A
\lor C) usando o sistema de tableaux analitico na linguagem
latex na biblioteca qtree”
```

Figura 10. Enunciado exemplo fornecido aos modelos

```
“ \Tree [.{T A \lor (B \land C) \\ F (A \lor B)\land (A \lor
C)} [.{F A \lor B} [.{ \color {blue} F A} \\ {\color{blue} F B}
[.{\color{blue}{T A}} [.{\times} ] ] [.{T B \land C} [.{{\color{blue} T
B} \\ T C} [.{\times} ] ] ] ] ] [.{F A \lor C} [.{{\color{blue} F A }
\\ {\color{blue} F C }} [.{\color{blue}{ T A }} [.{\times} ] ] [.{T B
\land C} [.{T B \\ {\color{blue} T C}} [.{\times} ] ] ] ] ] ] ”
```

Figura 11. Resposta exemplo fornecida aos modelos

```
“escreva a prova \vdash (A \lor (A \land B))\rightarrow A usando
o sistema de tableaux analítico na linguagem latex na biblioteca
qtree”
```

Figura 12. Pergunta fornecida aos modelos

5. Análise Experimental

Nesta etapa, avaliou-se se as respostas aplicavam corretamente as regras do Tableau Analítico e se continham erros lógicos ou incoerências estruturais.

5.1. Resultado das Perguntas Inseridas nos Modelos

As respostas obtidas podem ser acessadas através dos links: https://bit.ly/44fGrtS para as versões não-reflexivas e https://bit.ly/40k1wlr para as versões reflexivas. A Tabela 1 apresenta o desempenho dos modelos durante a resolução dos 40 exercícios.

Tabela 1. Desempenho Apresentado Pelos Modelos

Modelo	Questões Corretas	Erros Cometidos	Total de Acertos (%)
ChatGPT 4o	9	134	22,5%
ChatGPT 4o Reflexivo	6	123	13,6%
DeepSeek	18	80	45%
DeepSeek R1	16	48	40%

A Figura 13 apresenta uma prova corretamente resolvida por ambos os modelos e suas versões reflexivas para a questão $A \vee B, \neg B \vdash A$. Conforme o método os modelos marcam as premissas como verdadeiras ($T\,(A \vee B)$ e $T\,\neg B$) e negam a conclusão (FA). A partir dessas marcações, a expansão de $T \neg B$ gera FB, enquanto a expansão de $T(A \vee B)$ bifurca o tableau em dois ramos: um com TA, outro com TB. Em ambos os casos, surgem contradições, TA com FA em um ramo e TB com FB no outro, levando ao fechamento dos ramos e confirmando a validade do argumento.

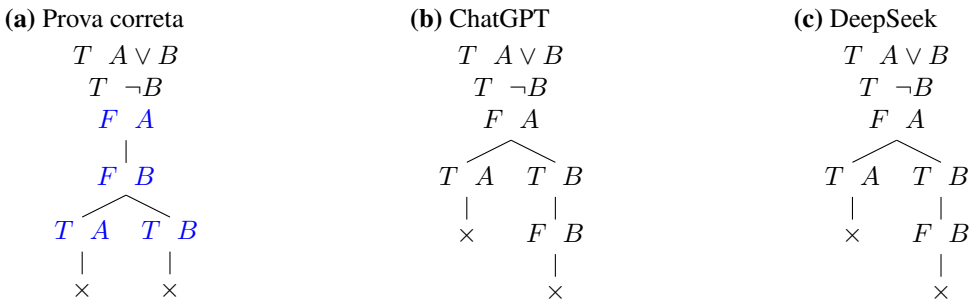

Figura 13. Comparação entre a prova correta e as respostas dos modelos para o enunciado $A \vee B, \neg B \vdash A$.

Embora o mecanismo reflexivo tenha reduzido o número total de erros, de 134 para 123 no ChatGPT e de 80 para 48 no DeepSeek, essa melhoria não resultou em mais respostas corretas. Nas versões padrão, os erros mais comuns incluíram expansão repetida de fórmulas, falhas na bifurcação de conectivos e atribuições incorretas de valores de verdade. Já nas versões reflexivas, surgiram novas inconsistências, como duplicação desnecessária de inferências e combinação inadequada de subfórmulas em um único ramo quando o correto seria bifurcar. No caso do ChatGPT, a versão reflexiva apresentou uma queda no número de acertos, de 9 para 6 questões. Isso indica que, embora a autocorreção ajude a mitigar certos erros, o processo reflexivo pode, em alguns casos, comprometer a fluidez do raciocínio, dificultando o encadeamento lógico da prova. A Figura 14 exemplifica uma dessas respostas incorretas geradas por ambas as versões do modelo.

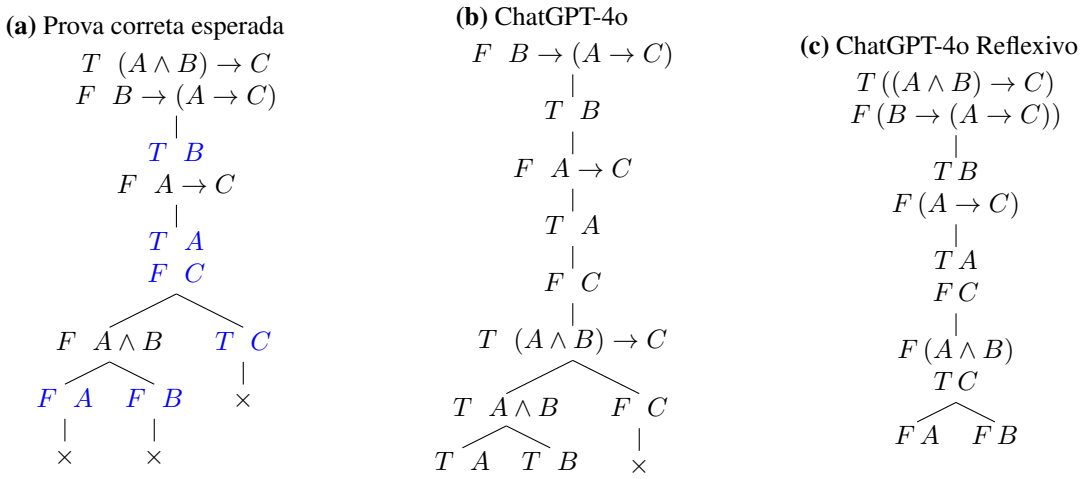

Figura 14. Comparação entre a prova correta e as respostas dos modelos para a questão 10: $(A \wedge B) \rightarrow C \vdash B \rightarrow (A \rightarrow C)$.

Na resposta da versão não reflexiva (Figura 14.b), foram identificados sete erros. O primeiro ocorre na aplicação inicial do tableau, com a fórmula $T\ (A \wedge B) \rightarrow C$ posicionada fora da estrutura adequada. Em seguida, na aplicação da implicação falsa, o modelo insere incorretamente $F\ A \rightarrow C$ em um novo ramo sob $T\ B$, quando ambas deveriam estar no mesmo ramo. Ao expandir $T\ (A \wedge B) \rightarrow C$, os valores de verdade são invertidos, e os ramos mais à esquerda permanecem abertos, enquanto o direito apresenta uma contradição sem justificativa. Já na versão reflexiva (Figura 14.c), observa-se uma redução para três erros. No entanto, persiste a falha na aplicação da implicação verdadeira, com

subfórmulas colocadas no mesmo ramo em vez de bifurcar. Além disso, os ramos não são devidamente fechados, e a contradição, embora presente, não é formalmente expressa.

No caso dos modelos DeepSeek e DeepSeek R1, a versão não reflexiva já apresentava desempenho consistente, com 18 questões corretas e 80 erros. A versão reflexiva manteve um bom nível de acertos (16), mas reduziu significativamente o número de erros para 48. Isso sugere que, em modelos com boa aderência às regras do Tableau, o mecanismo reflexivo pode ser útil para reduzir inconsistências sem comprometer o desempenho geral. A Figura 15 exemplifica uma resposta incorreta gerada por esse modelo.

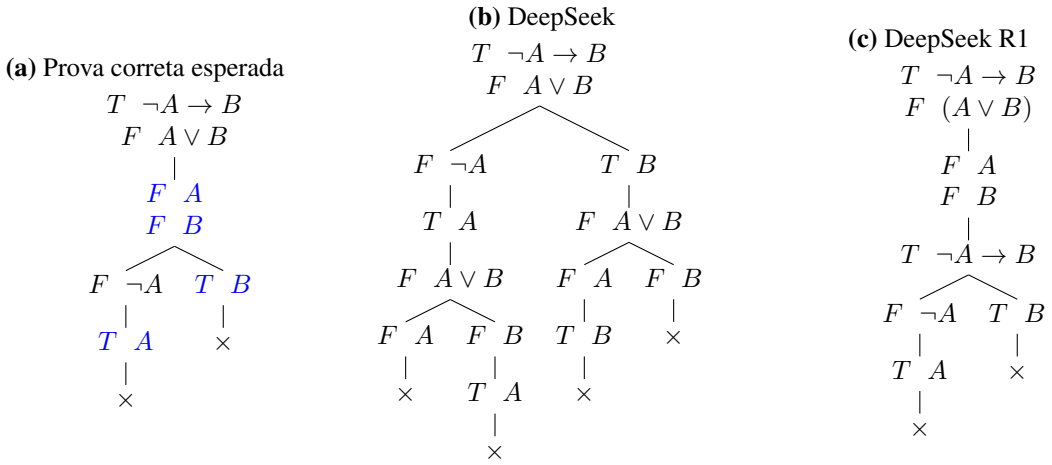

Figura 15. Comparação entre a prova correta e as respostas do DeepSeek R1 para a questão 31: $\neg A \rightarrow B \vdash A \vee B$.

Na resposta analisada (Figura 15.b), o modelo cometeu seis erros. O principal foi a repetição indevida da fórmula $F\ A \vee B$, violando a regra de expansão única. Além disso, $T\ A$ e $T\ B$ foram expandidos redundante e incorretamente em ramos distintos. O modelo também aplicou de forma equivocada a regra da disjunção falsa, tratando $F\ A \vee B$ como bifurcação em vez de manter as subfórmulas no mesmo ramo — erro que se repetiu em diferentes partes do tableau. Na versão reflexiva (Figura 15.c), o número de erros foi reduzido para apenas um: a repetição da fórmula $T\ \neg A \rightarrow B$, novamente em desacordo com a regra de expansão única.

6. Conclusão e Trabalhos Futuros

Este estudo avaliou o desempenho dos modelos ChatGPT-4o e DeepSeek, com e sem mecanismos reflexivos, na resolução de provas de Lógica Proposicional utilizando o método Tableau Analítico. Os resultados indicam que, embora os modelos apresentem algum domínio das regras básicas do sistema, ambos ainda cometem erros frequentes, como expansões incorretas, bifurcações inválidas e falhas no fechamento dos ramos. A introdução do mecanismo reflexivo resultou em uma redução geral no número de erros, especialmente no DeepSeek R1. No entanto, essa melhoria estrutural não se traduziu em mais acertos completos, como observado em ambos os modelos, cuja versão reflexiva teve desempenho inferior ao padrão em número de acertos. Isso sugere que a autocorreção pode interferir na continuidade do raciocínio.

Diante dessas limitações e considerando a constante evolução dos LLMs, há diversas possibilidades para estudos futuros que podem expandir e melhorar essa análise. Uma

alternativa seria ampliar o conjunto de questões utilizadas nos testes, incluindo desafios de maior complexidade, além de explorar estratégias de treinamento como o fine-tuning. O uso de outros LLMs, também pode enriquecer significativamente os resultados.

Agradecimentos

Agradecemos ao Kunumi Lab UFC, financiado pelo Instituto Kunumi, pelo apoio financeiro para realização desta pesquisa.

Referências

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Aydin, O., Karaarslan, E., Erenay, F. S., and Bacanin, N. (2025). Generative ai in academic writing: A comparison of deepseek, qwen, chatgpt, gemini, llama, mistral, and gemma. *arXiv preprint arXiv:2503.04765*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Carbonell, J. G., Michalski, R. S., and Mitchell, T. M. (1983). An overview of machine learning. *Machine learning*, pages 3–23.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Huth, M. (2004). *Logic in Computer Science: Modelling and Reasoning about Systems*. Cambridge University Press.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Koubaa, A. (2023). Gpt-4 vs. gpt-3.5: A concise showdown.

Lalwani, A., Chopra, L., Hahn, C., Trippel, C., Jin, Z., and Sachan, M. (2024). Nl2fol: Translating natural language to first-order logic for logical fallacy detection. *arXiv preprint arXiv:2405.02318*.

Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D., et al. (2024). Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., and Zhang, Y. (2023). Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

Martins, F. L. B., Oliveira, A. C. A., Vasconcelos, D. R., and de Menezes, M. V. (2025). Avaliando a habilidade do chatgpt de realizar provas de dedução natural em lógica proposicional e lógica de predicados. *Revista Brasileira de Informática na Educação*, 33:244–278.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

OpenAI (2021). ChatGPT. `https://openai.com/research/chatgpt`. Acesso em: 03 de agosto de 2024.

Saparov, A., Pang, R. Y., Padmakumar, V., Joshi, N., Kazemi, M., Kim, N., and He, H. (2023). Testing the general deductive reasoning capacity of large language models using ood examples. *Advances in Neural Information Processing Systems*, 36:3083–3105.

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., and Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Siskind, J. M. and Dimitriadis, A. (2008). Qtree, a latex tree-drawing package.

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., and Agyemang, B. (2023). What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education. *Smart learning environments*, 10(1):15.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vasconcelos, D. R. (2023). Anita: Analytic tableau proof assistant. *arXiv preprint arXiv:2303.05864*.

Viegas, C. V. et al. (2024). Avaliando a capacidade de llms na resolução de questões do poscomp.

Zhang, M. and Li, J. (2021). A commentary of gpt-3 in mit technology review 2021. *Fundamental Research*, 1(6):831–833.