CBIE 2023
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO
Uma escola para o futuro: Tecnologia e conectividade a serviço da educação

Avaliando a habilidade do ChatGPT de realizar provas de Dedução Natural em Lógica Proposicional

Francisco Leonardo Batista Martins[1] **, Augusto César Araújo de Oliveira** [1]
Davi Romero de Vasconcelos , Maria Viviane de Menezes [1]

[1]Universidade Federal do Ceará - Campus Quixadá

{lmartins,augustces}@alu.ufc.br

{vivianemenezes,daviromero}@ufc.br

Abstract. *The use of conversational agents (chatbots) in education has sparked growing interest among researchers, educators, and educational institutions. These systems have the ability to comprehend and process large quantity of data, offering individualized support to students. However, it is important to consider that they can also generate incorrect responses in some tasks: such as logical reasoning. This paper aims to evaluate the ability of the conversational agent ChatGPT to solve Natural Deduction exercises in propositional logic. The study seeks to determine whether ChatGPT is a suitable tool for this task. To achieve this, experiments are conducted using a database of exercises in Natural Deduction. This study aims to contribute to the understanding of the capabilities and limitations of conversational agents in logical reasoning skills.*

Resumo. *A utilização de agentes conversacionais, também conhecidos como chatbots, na educação tem despertado um crescente interesse de pesquisadores, educadores e instituições de ensino em todo o mundo. Esses sistemas têm a capacidade de compreender e processar grandes volumes de dados, oferecendo suporte individualizado aos alunos. No entanto, é importante considerar que esses sistemas podem gerar respostas incorretas em tarefas que envolvem raciocínio lógico. Este artigo tem como objetivo avaliar a habilidade do agente conversacional ChatGPT na resolução de exercícios de Dedução Natural em lógica proposicional. O estudo busca verificar se o ChatGPT é uma ferramenta adequada para essa tarefa. Para isso, são realizados experimentos utilizando uma base de dados de exercícios de dedução natural em lógica proposicional. Esse estudo busca contribuir para a compreensão das capacidades e limitações dos agentes conversacionais em habilidades de raciocínio lógico.*

1. Introdução

A utilização de *agentes conversacionais*, também conhecidos como *chatbots*, na educação tem despertado um crescente interesse de pesquisadores, educadores e instituições de ensino em todo o mundo [Weber et al. 2021, Tlili et al. 2023, Kasneci et al. 2023]. A capacidade desses sistemas em compreender e processar grandes volumes de dados, além de sua habilidade em aprender e adaptar-se a novas informações, oferece oportunidades promissoras para aprimorar o processo de ensino-aprendizagem. Um dos principais benefícios de se utilizar tais ferramentas é a possibilidade de oferecer suporte individualizado aos alunos. No entanto, é importante estar ciente de que esses sistemas podem gerar

respostas incorretas em determinadas situações, especialmente quando lidam com tarefas envolvendo *raciocínio lógico* [Liu et al. 2023].

A Lógica para Computação é uma disciplina abordada em grande parte dos cursos de graduação na área de Tecnologia da Informação e Comunicação (TICs). O objetivo da lógica na computação é desenvolver linguagens para modelar as situações do mundo e dos sistemas, de modo que possamos analisá-las formalmente, construindo argumentos sobre elas para serem apresentados e *justificados rigorosamente*. A Dedução Natural é um dos conteúdos mais importantes na ementa desta disciplina, sendo utilizada para derivar *conclusões* a partir de sentenças dadas como verdadeiras (as quais são denominadas *premissas*), seguindo regras específicas [Pelletier 1999, Huth and Ryan 2004].

Neste contexto, este artigo tem como objetivo avaliar a habilidade do agente conversacional ChatGPT, um modelo de linguagem desenvolvido pela OpenAI [OpenAI 2021], na resolução de exercícios de Dedução Natural em lógica proposicional. A principal pergunta que pretendemos responder com esse estudo é: *o ChatGPT é confiável na tarefa solucionar exercícios de dedução natural em lógica proposicional?* Para isso, avaliamos o desempenho da ferramenta em um base de dados de exercícios de dedução natural em lógica proposicional, disponibilizadas para alunos de graduação da disciplina de Lógica para Computação. As iterações com o modelo foram feitas utilizando uma API (*Application Programming Interface*) disponibilizada pela OpenAI acessada por meio de sua biblioteca na linguagem de programação Python. Por fim, o desempenho do modelo na demonstração dos exercícios foi avaliado verificando-se a correção das respostas e, no caso de apresentação respostas incorretas, avaliamos se os erros eram erros de aplicação de regras lógicas ou apenas erros de escrita da prova.

O restante do artigo está organizado como mostrado a seguir. Na Seção 2, apresentaremos a fundamentação teórica do trabalho. Na Seção 3, apontamos alguns trabalhos relacionados sobre o uso do ChatGPT em tarefas de raciocínio lógico. Na Seção 4, descrevemos a metodologia utilizada para avaliar a ferramenta. Na Seção 5, apresentamos os resultados obtidos. E, na Seção 6 apresentaremos as conclusões e trabalhos futuros.

2. Fundamentação

2.1. Lógica Proposicional

A Lógica Proposicional se baseia em *proposições* ou *frases declarativas* que se pode argumentar sobre sua veracidade ou falsidade. As frases declarativas são consideradas como *atômicas* (e.g, "*o dólar sobe*", "*os produtos ficam mais caros*") de certa forma que podemos atribuir símbolos distintos $P, Q, R, \cdots$ a cada uma destas frases (e.g, P : "*o dólar sobe*", Q : "*os produtos ficam mais caros*"). Para codificar frases mais complexas usamos os conectivos lógicos: $\neg$ (negação), $\wedge$ (conjunção), $\vee$ (disjunção) e $\rightarrow$ (implicação) [Enderton 1972]. Por exemplo: $P \wedge Q$ codifica a frase "O dólar sobe *e* os produtos ficam mais caros"; $P \vee Q$ codifica a frase "O dólar sobe *ou* os produtos ficam mais caros."; $P \rightarrow Q$ representa "*se* O dólar sobe *então* os produtos ficam mais caros." e; $\neg P$: "O dólar não subiu".

O conjunto dos átomos proposicionais, juntamente com os conectivos e os símbolos '(', ')' formam o alfabeto da linguagem da Lógica Proposicional. A sintaxe da Lógica Proposicional pode ser definida por uma gramática na forma de *Backus Naur*

(BNF - *Backus Naur Form*) como a seguir:

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

em que $\bot$ e P representam, respectivamente, a contradição e qualquer proposição atômica e cada ocorrência φ a direta de '::=' representa qualquer fórmula já construída.

O sistema de Dedução Natural é um mecanismo que permite a construção de uma prova formal, estabelecendo uma conclusão φ a partir de um conjunto de premissas $\Gamma = \{\varphi_1, \varphi_2, \cdots, \varphi_n\}$, denotado por $\Gamma \vdash \varphi$, aplicando-se sucessivamente *regras de demonstração*. O artigo [Pelletier 2000] compara o estilo de demonstrações de Dedução Natural de 33 livros-texto, sendo o estilo *Fitch* (caixas) o adotado na maioria deles.

Neste estilo, as demonstrações são apresentadas de forma linear e sequencial, na qual cada uma das linhas da prova é numerada, tem uma afirmação e uma justificativa. As justificativas são definidas por serem *premissas* da prova ou pela aplicação de uma das *regras do sistema dedutivo*. As subprovas dentro de uma prova maior têm *caixas* ao redor e servem para delimitar o escopo de hipóteses temporárias. Provas podem ter caixas dentro de caixas, ou pode-se abrir outras caixas depois de fechar outras, obedecendo as regras de demonstração. Uma fórmula só pode ser utilizada em uma prova em um determinado ponto se essa fórmula aconteceu anteriormente e se nenhuma caixa que contenha essa ocorrência da fórmula tenha sido fechada.

Uma demonstração em Dedução Natural é construída a partir da enumeração das premissas e da aplicação das *regras de Dedução Natural*. Em geral, cada conectivo possui uma regra para adicionar e outra regra para eliminar este conectivo. A construção de uma demonstração é um exercício criativo. Não é óbvio quais regras aplicar e em qual ordem, a partir das premissas, para obter a conclusão desejada. A seguir, serão apresentadas algumas regras de Dedução Natural.

Premissas O primeiro passo em uma demonstração em Dedução Natural no estilo *Fitch* é enumerar as premissas da prova. A Figura 1 apresenta a estrutura geral da regra das premissas, na qual $\varphi_1, \varphi_2, \ldots, \varphi_n$ são representadas em uma linha cada, seguindo uma numeração sequencial e como justificativa "premissa".

1.	φ_1	premissa
2.	φ_2	premissa
$\vdots$	$\vdots$	$\vdots$
n.	φ_n	premissa
$\vdots$	$\vdots$	$\vdots$

Figura 1. Enumeração das premissas.

Conjunção Introdução ($\wedge i$) A regra da introdução da conjunção ($\wedge i$) é apresentada na Figura 2a (ou Figura 2b), na qual a fórmula $\varphi \wedge \psi$ pode ser concluída em uma linha p se φ e ψ foram demonstradas nas linhas m (ou n) e n (ou m), respectivamente, anteriores a linha p e que não foram descartadas. A Figura 2c exibe a aplicação $\wedge i$ da fórmula $A \wedge B$ na linha 3 a partir das fórmulas A e B, definidas nas linhas 1 e 2, respectivamente, que são anteriores a linha 3.

$\vdots$	$\vdots$	$\vdots$
$m.$	φ	
$\vdots$	$\vdots$	$\vdots$
$n.$	ψ	
$\vdots$	$\vdots$	$\vdots$
$p.$	$\varphi \wedge \psi$	$\wedge$i m,n

(a) Regra $\wedge i$

$\vdots$	$\vdots$	$\vdots$
$m.$	ψ	
$\vdots$	$\vdots$	$\vdots$
$n.$	φ	
$\vdots$	$\vdots$	$\vdots$
$p.$	$\varphi \wedge \psi$	$\wedge$i m,n

(b) Regra $\wedge i$

1.	A	premissa
2.	B	premissa
3.	$A \wedge B$	$\wedge i$ 1,2

(c) $A, B \vdash A \wedge B$

Figura 2. Conjunção Introdução ($\wedge i$)

Conjunção Eliminação ($\wedge e$) A regra da eliminação da conjunção ($\wedge e$) é apresentada na Figura 3a (ou Figura 3b), na qual a fórmula φ (ou ψ) pode ser concluída na linha p a partir da eliminação à esquerda (ou à direita) da conjunção da fórmula $\varphi \wedge \psi$ da linha n (anterior a m e não foi descartada). A Figura 3c exibe uma aplicação da regra na qual A é obtida na linha 3 pela eliminação da conjunção à esquerda da fórmula $A \wedge B$ da linha 1.

$\vdots$	$\vdots$	$\vdots$
$m.$	$\varphi \wedge \psi$	
$\vdots$	$\vdots$	$\vdots$
$p.$	φ	$\wedge$e m

(a) Regra $\wedge e$ à esquerda

$\vdots$	$\vdots$	$\vdots$
$m.$	$\varphi \wedge \psi$	
$\vdots$	$\vdots$	$\vdots$
$p.$	ψ	$\wedge$e m

(b) Regra $\wedge e$ à direita

1.	$A \wedge B$	premissa
2.	C	premissa
3.	A	$\wedge e$ 1
4.	$A \wedge C$	$\wedge i$ 3,2

(c) $A \wedge B, C \vdash A \wedge C$

Figura 3. Conjunção Eliminação ($\wedge e$)

Implicação Eliminação ($\rightarrow$e) A regra da eliminação da implicação ($\rightarrow$e), também conhecida como *Modus Ponens*, é apresentada na Figura 4a (ou Figura 4b), na qual a fórmula ψ pode ser concluída na linha p a partir da eliminação da implicação da fórmula $\varphi \rightarrow \psi$ da linha m (ou n) e φ da linha n (ou m), anteriores a p e não descartadas. A Figura 5b exibe uma aplicação da regra na qual a fórmula C é obtida na linha 4 pela eliminação da implicação das fórmulas $B \rightarrow C$ e B das linha 2 e 3, respectivamente.

$\vdots$	$\vdots$	$\vdots$
$m.$	$\varphi \rightarrow \psi$	
$\vdots$	$\vdots$	$\vdots$
$n.$	φ	
$\vdots$	$\vdots$	$\vdots$
$p.$	ψ	$\rightarrow$e m, n

(a) Regra $\rightarrow e$

$\vdots$	$\vdots$	$\vdots$
$m.$	φ	
$\vdots$	$\vdots$	$\vdots$
$n.$	$\varphi \rightarrow \psi$	
$\vdots$	$\vdots$	$\vdots$
$p.$	ψ	$\rightarrow$e m, n

(b) Regra $\rightarrow e$

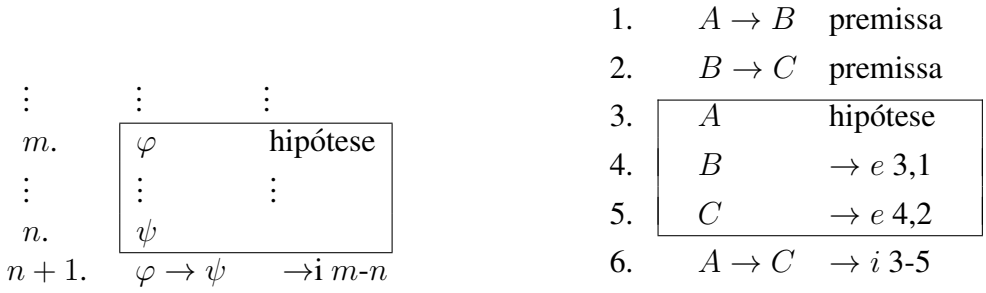
1.	$A \wedge B$	premissa
2.	$B \rightarrow C$	premissa
3.	B	$\wedge e$ 1
4.	C	$\rightarrow e$ 2,3

(c) $A \wedge B, B \rightarrow C \vdash C$

Figura 4. Implicação Eliminação ($\rightarrow e$)

Implicação Introdução ($\rightarrow$i) A regra da introdução da implicação ($\rightarrow$i) constrói condicionais que não ocorrem como premissas. Para construção de um condicional é necessário

realizar *raciocínio hipotético*, isto é, supor *temporariamente* que uma dada fórmula é verdadeira. Chamamos esta fórmula de *hipótese*. Assim, utilizamos *caixas de demonstração*, que servem para delimitar o *escopo da hipótese temporária*. Observe na Figura 5a que para provar o condicional $\varphi \to \psi$ na linha $n+1$, colocamos φ como *hipótese* no topo de uma caixa (linha m), aplicamos um número finito de regras de forma a obter ψ na linha n. Todo o raciocínio para obter ψ depende da veracidade de φ e, por isso, as fórmulas resultante deste raciocínio ficam delimitadas na caixa. Na linha seguinte $(n+1)$ podemos aplicar a regra $\to i$ para obter $\varphi \to \psi$, sendo que este condicional não mais depende da hipótese φ. Na justificativa da linha $n+1$ utilizamos o nome da regra seguido das linhas inicial e final da caixa ($\to i$ m-n). A Figura 5b exibe uma aplicação da regra na qual a fórmula $A \to C$ é obtida na linha 6 a partir da caixa das linhas 3 a 5 em que A é a hipótese e C é a conclusão da caixa.



(a) Regra $\to i$ **(b)** $A \to B, B \to C \vdash A \to C$

Figura 5. Implicação Introdução ($\to i$)

Demonstrações podem ter caixas dentro de caixas, ou pode-se abrir novas caixas depois de fechar outras. No entanto, existem regras sobre quais fórmulas podem ser utilizadas em que ponto na demonstração. Em geral, só podemos usar uma fórmula em um determinado ponto se esta fórmula ocorre *antes* desse ponto e se nenhuma caixa que contenha a ocorrência desta fórmula tenha sido fechada [Huth and Ryan 2004].

Por limitações de espaço, as definições das demais regras não serão detalhas neste artigo, quais sejam: Negação Introdução ($\neg i$); Negação Eliminação ($\neg e$); Disjunção Introdução ($\vee i$); Disjunção Eliminação ($\vee e$); Redução Ao Absurdo (raa); Contradição Eliminação ($\bot e$); e Copie. Para maiores detalhes, consulte [Huth and Ryan 2004].

2.2. Redes Neurais Transformacionais

Aprendizado de Máquina é uma subárea da Inteligência Artificial que estuda modelos e algoritmos capazes de aprender a partir de dados. As *Redes Neurais Artificiais* (RNAs) fundamentam-se na estrutura e funcionamento do cérebro humano, consistindo em um conjunto interconectado de unidades de processamento chamadas neurônios artificiais, que são organizados em camadas. Cada neurônio recebe entradas, aplica uma função de ativação e gera uma saída. As informações são passadas por meio das camadas até que a saída seja gerada. As RNAs são usadas em uma ampla variedade de aplicações, como reconhecimento de padrões, classificação, regressão, visão computacional e processamento de linguagem natural [Russell 2010].

A arquitetura da rede neural está intimamente relacionada ao tipo de problema que se pretende resolver: certas arquiteturas são mais adequadas para certos tipos de dados e

tarefas. Como exemplos de arquiteturas de redes neurais temos: *Multilayer Perceptrons* (MLPs), CNNs (Redes Neurais Convolucionais) e RNNs (Redes Neurais Recorrentes).

As Redes Neurais Transformacionais (ou simplesmente *Transformers*) [Vaswani et al. 2017] são uma arquitetura de redes neurais que têm desempenhado um papel fundamental no Processamento de Linguagem Natural (PLN). Utilizam mecanismos de *atenção* para permitir que cada elemento de entrada se comunique diretamente com todos os outros elementos da rede, o que fornece a capacidade de capturar relações de longo alcance nas sequências e lidar com contextos complexos.

O ChatGPT [OpenAI 2021] é uma tecnologia de chatbots desenvolvida pela OpenAI que utiliza a rede neural transformacional *Generative Pre-trained Transformer* (GPT-3) [Brown et al. 2020]. Ele foi treinado em uma gigantesca quantidade de dados, compreendendo bilhões de palavras e trechos de texto de várias fontes na internet. Esse treinamento permitiu que a ferramenta aprendesse a gramática, o estilo e o contexto das diferentes formas de comunicação humana. Uma das principais características do GPT-3 é a sua capacidade de gerar texto de forma autônoma. Ao receber uma pergunta ou uma instrução, o modelo processa a informação, consulta seu vasto conhecimento prévio e gera uma resposta coerente e relevante. Ele pode incorporar o contexto da conversa anterior para fornecer respostas mais personalizadas e contextualmente apropriadas.

No entanto, é importante ressaltar que o GPT-3 é um modelo de linguagem estatístico e não possui compreensão profunda sobre certos tópicos. Embora seja capaz de produzir respostas impressionantes em muitas situações, também pode gerar respostas incorretas. A confiabilidade das respostas do ChatGPT depende da qualidade e da precisão dos dados com os quais o modelo foi treinado.

3. Trabalhos Relacionados

A literatura sobre o uso de grandes modelos generativos na tarefa de solucionar exercícios envolvendo conteúdo de lógica é bem recente, sendo a maioria dos trabalhos preliminares. Nesta seção, apresentamos 3 trabalhos que são relacionados com a proposta deste artigo.

O trabalho de [Liu et al. 2023] avaliou a habilidade de raciocínio lógico do ChatGPT em domínios *benchmarks* tais como LogiQA [Liu et al. 2020] e ReClor [Yu et al. 2020]. As perguntas realizadas ao *chat* foram de dois tipos: Questões de múltipla escolha com quatro itens e; Questões de verificar se uma conclusão segue logicamente de premissas (Figura 6) nas quais o *chat* deve responder *sim*, *não* ou *não é possível determinar*. Os autores mostram que ChatGPT tem um bom desempenho ao responder as perguntas dos domínios *benchmarks*. No entanto, o desempenho da ferramenta não é tão bom ao responder perguntas de novos conjuntos de dados.

O trabalho [Carl 2023] tem o objetivo de avaliar o desempenho do modelo text-davinci-003 (*text completion model*) [Ye et al. 2023], incluído na arquitetura do GPT, nas tarefas de *correção automática* de exercícios envolvendo tradução de sentenças em linguagem natural para fórmulas lógicas (formalização) e de fórmulas lógicas para sentenças em linguagem natural (desformalização) e; (ii) exercícios de argumentação em linguagem natural envolvendo raciocínio lógico. Nos exercícios de *formalização* é dada uma frase em linguagem natural e o aluno deve produzir como resposta uma fórmula lógica que corresponda a essa frase, conforme mostrado na Figura 7. Os resultados deste trabalho

Premise: Ten new television shows appeared during the month of September. Five of the shows were sitcoms, three were hourlong dramas, and two were news-magazine shows. By January, only seven of these new shows were still on the air. Five of the shows that remained were sitcoms.
Conclusion: At least one of the shows that were cancelled was an hourlong drama.

Figura 6. Exemplo de exercícios para verificar se uma conclusão segue das premissas [Liu et al. 2023].

são bem preliminares, no entanto, é um rascunho de avaliação de desempenho de modelos generativos em tarefas envolvendo especificação de fórmulas e raciocínio em lógica.

Seja S a declaração *O sol brilha.* e W a declaração *Eu saio para passear*. Formalize a seguinte frase *Não vou sair para passear a menos que o sol brilhe* em lógica proposicional.

Figura 7. Exemplo de exercício de formalização [Carl 2023].

Em [vom Scheidt 2023], os autores avaliaram o desempenho do GPT-4 em tarefas matemáticas tais como: declaração de funções, provas de teoremas e verificação de provas fornecidas pelo usuário. Para isso, foi utilizada uma linguagem formal (não publicada), denominada *Axiotome*. Foram fornecidas a sintaxe e semântica desta linguagem para a interface gráfica do ChatGPT e, em seguida, foram realizadas as seguintes perguntas [1]: (i) defina uma função disjunção em *Axiotome*; (ii) Defina uma função ternária "*se*"; (iii) Prove o teorema referente à primeira Lei de De Morgan $\neg(A \vee B) \vdash \neg A \wedge \neg B$; verifique a correção da prova de $\neg\neg A \vdash A$ e verifique a correção da prova de $A \vee B \vdash B \vee A$. O sistema completou as tarefas de definição de função e verificação de provas de teoremas com sucesso (tarefas i, ii, iv e v). No entanto, na tarefa de provar um teorema (tarefa iii), o sistema apresentou alguns erros na prova.

4. Metodologia

Para avaliar o desempenho do modelo GPT-3.5-turbo[2] na tarefa de construir provas de exercícios de Dedução Natural em Lógica Proposicional foram utilizados um conjunto de 41 exercícios aplicados durante a disciplina de Lógica para Computação para alunos de graduação em cursos da área de Tecnologia da Informação[3]. Os exercícios são textos em LaTeX, como exemplificado a seguir: `$A \rightarrow B, \lnot B \vdash \lnot A$`, o que corresponde ao teorema em lógica proposicional $A \rightarrow B, \neg B \vdash \neg A$ com premissas $A \rightarrow B$ e $\neg B$ e conclusão $\neg A$.

Inicialmente, um exercício exemplo foi submetido à ferramenta cujo enunciado é um texto com um teorema em LaTeX (conforme mostrado na Figura 8) e a saída é um código LaTeX (conforme mostrado na Figura 9), no formato adotado pelo livro texto "*Lógica em Ciência da Computação: Modelagem e Raciocínio sobre sistemas*"

[1] Iremos reportar aqui apenas as perguntas relacionadas com lógica proposicional
[2] `https://platform.openai.com/docs/models`
[3] `https://github.com/leonardomartins777/ChatGPT_Naturalded`

[Huth and Ryan 2004], o qual utiliza o pacote `logicproof`[4] de escrita de provas de dedução natural no estilo *Fitch*. Em seguida, novos exercícios são submetidos ao modelo para que ele construa a sua solução, como o exemplo apresentado na Figura 10.

"escreva a prova `$A \rightarrow B, B \rightarrow C \vdash A \rightarrow C$` usando o sistema de dedução natural no estilo fitch e o pacote latex logicproof."

Figura 8. Pergunta exemplo fornecida ao modelo.

```
\begin{logicproof}{6}
    A \rightarrow B & premissa \\
    B \rightarrow C & premissa \\
    \begin{subproof}
        A & hipótese \\
        B & $\rightarrow e$ 3,1 \\
        C & $\rightarrow e$ 4,2
    \end{subproof}
    A \rightarrow C & $\rightarrow i$ 3-5
\end{logicproof}
```

Figura 9. Resposta exemplo fornecida ao modelo.

"escreva a prova `\vdash A \rightarrow (B \rightarrow A)` usando o sistema de dedução natural no estilo fitch e o pacote latex logicproof"

Figura 10. Pergunta fornecida para que o modelo construa sua demonstração.

Para a comunicação com o modelo *GPT-3.5-turbo* foi utilizada a API (do inglês, *Application Programming Interface*) disponibilizada pela *OpenAI*[5], que foi acessada por meio de sua biblioteca na linguagem de programação *Python*. O modelo recebe uma conversa, que consiste em uma lista de mensagens como entrada e retorna uma mensagem gerada como saída.

5. Análise Experimental

Avaliamos se nas respostas fornecidas *as regras de Dedução Natural eram aplicadas corretamente* e se as provas foram escritas de acordo com a sintaxe do pacote `logicproof`.

A Figura 11 mostra um exemplo de demonstração correta para $\vdash (A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$. Nesta resposta, o modelo escreveu um código LaTeX, o qual foi compilado e corresponde a demonstração mostrada na figura. Veja em mais detalhes que o modelo: fez a suposição correta das hipóteses nas linhas 1, 2 e 3; aplicou corretamente a regra da implicação eliminação ($\rightarrow e$) e referenciou corretamente as linhas com a fórmula contendo o condicional e a veracidade do antecedente nas linhas 4, 5 e 6; aplicou corretamente a regra da implicação introdução ($\rightarrow i$) e referenciou corretamente os intervalos das caixas nas linhas 7, 8 e 9.

[4] `https://ctan.org/pkg/logicproof`

[5] `https://platform.openai.com/docs/api-reference`

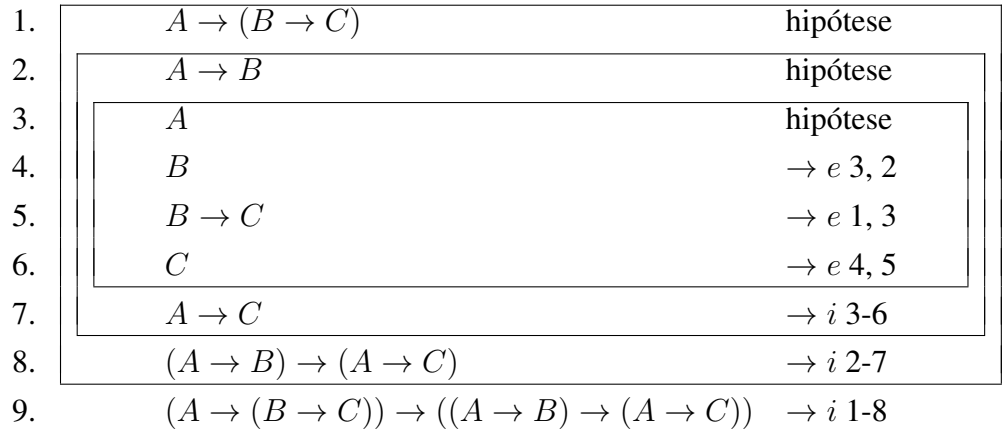


Figura 11. Prova correta para $\vdash (A \to (B \to C)) \to ((A \to B) \to (A \to C))$, resultado da compilação do código LaTeX escrito pelo modelo.

No entanto, a maior parte das respostas obtidas continham algum tipo de erro, os quais foram classificados em: *erros lógicos*, quando ocorrem erros na aplicação das regras de Dedução Natural e; *erros de referências nas regras*, quando as respostas, apesar das regras serem utilizadas adequadamente, apresentavam erros nas referências durante sua aplicação.

Nos *erros lógicos* avaliamos todos os erros decorrentes do mal uso das regras de Dedução Natural e incoerências lógicas em geral, que são os erros mais críticos para o objetivo do experimento. A partir dos resultados podemos observar que a maioria das respostas apresentam erros lógicos, um total de 31 (75,61%). Os erros de escrita de prova foram bem frequentes nestes experimentos. Destacamos aqui as questões que tiveram exclusivamente erros de referência de linhas, no qual são feitas referências incorretas às linhas durante a aplicação de determinadas regras. Esses erros totalizaram apenas 04 questões (9,76%). Dessa forma, podemos observar uma grande imprecisão do modelo na tarefa de escrever as demonstrações, totalizando 35 questões erradas (85,37%). Em algumas respostas produzidas pelo modelo, há vários erros dessa categoria. A Figura 12 ilustra uma resposta incorreta produzida pelo modelo na qual há erros de escrita de prova (referência de linhas) e também erros lógicos. Nesta demonstração, o modelo: referenciou incorretamente nas linhas 3 e 4, as linhas $1, 1$, uma vez que a aplicação da regra $\vee\, e$ deveria referenciar apenas a linha 2; referenciou incorretamente na linha 6, a linha 2 na aplicação da regra $\wedge\, e$, uma vez que deveria ter mencionado a linha 5 (o mesmo ocorre na linha 8); referenciou incorretamente na linha 7, as linhas $3, 2$, uma vez que a aplicação da regra $\vee$ deveria referenciar apenas a linha 6; referenciou incorretamente na linha 9, as linhas $3, 4$, uma vez que a aplicação da regra $\vee$ deveria referenciar apenas a linha 8 e; **cometeu um erro lógico** na linha 10 no uso da regra $\wedge$ i, a qual não poderia ser aplicada nesse contexto. Conforme a prova foi escrita, o correto seria utilizar a regra $\vee$ e para obter $A \vee C$. No entanto, não seria possível obter $(A \vee B)$.

A Tabela 1 resume o desempenho do modelo ChatGPT na tarefa de Dedução Natural, onde ele teve apenas 06 acertos completos num total de 41 exercícios (14,63%). Além disso, podemos observar os tipos de erros mais comuns apresentados pelo modelo, sendo que para cada resposta mais de um erro pode ter ocorrido.

1.	$A \vee (B \wedge C)$	premissa
2.	A	hipótese
3.	$A \vee B$	$\vee i$ 1, 1
4.	$A \vee C$	$\vee i$ 1, 1
5.	$B \wedge C$	hipótese
6.	B	$\wedge e_1$ 2
7.	$A \vee B$	$\vee i$ 3, 2
8.	C	$\wedge e_2$ 2
9.	$A \vee C$	$\vee i$ 3, 4
10.	$(A \vee B) \wedge (A \vee C)$	$\wedge i$ 2-6

Figura 12. Demonstração incorreta para $A \vee (B \wedge C) \vdash (A \vee B) \wedge (A \vee C)$, resultado da compilação do código LaTeX escrito pelo modelo.

Tabela 1. Desempenho apresentado pelo modelo na tarefa de dedução natural

Modelo	Acertos	Erros Apenas de Referência	Erros Lógicos
ChatGPT	6	4	31
Total (%)	14,63%	9,76%	75,61%

6. Conclusões e Trabalhos Futuros

Neste artigo, avaliamos o desempenho do ChatGPT na tarefa de construir provas em dedução natural na lógica proposicional. Após a análise dos resultados, fica evidente que o modelo ainda apresenta limitações significativas nesse domínio. Dos 41 problemas fornecidos para o ChatGPT, o modelo conseguiu acertar apenas 06 exercícios. Essa taxa de acerto relativamente baixa indica que o ChatGPT ainda não possui a capacidade completa de entender e raciocinar corretamente sobre a lógica proposicional.

Várias são as razões possíveis para o desempenho insatisfatório do modelo. Em primeiro lugar, a construção de provas em lógica proposicional requer um conhecimento profundo sobre as regras de inferência. O ChatGPT pareceu ter dificuldades tanto em selecionar uma regra apropriada como também em aplicar corretamente essas regras. Além disso, é importante considerar que o modelo foi treinado em uma ampla variedade de textos e não recebeu treinamento específico em lógica proposicional.

Como trabalhos futuros, pretendemos: treinar o modelo com todas as regras de dedução natural em lógica proposicional para verificar se há melhora no desempenho; utilizar bancos de questões públicos como *dataset*; verificar o desempenho do modelo na tarefa de construir provas de dedução natural em lógica de predicados e; verificar o desempenho do modelo na tarefa de resolução de exercícios de outras temáticas envolvendo lógica proposicional e lógica de predicados.

Referências

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.

Carl, M. (2023). Using large language models for (de-) formalization and natural argumentation exercises for beginner's students. *arXiv preprint arXiv:2304.06186*.

Enderton, H. B. (1972). *A mathematical introduction to logic*. Academic Press.

Huth, M. and Ryan, M. (2004). *Logic in Computer Science: Modelling and Reasoning about Systems (2nd Ed.)*. 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.

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*.

Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., and Zhang, Y. (2020). Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

OpenAI (2021). ChatGPT. `https://openai.com/research/chatgpt`. Acesso em: 13 de junho de 2023.

Pelletier, F. J. (1999). A brief history of natural deduction. *History and Philosophy of Logic*, 20(1):1–31.

Pelletier, F. J. (2000). A history of natural deduction and elementary logic textbooks. *Logical consequence: Rival approaches*, 1:105–138.

Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.

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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

vom Scheidt, G. (2023). Experimental results from applying gpt-4 to an unpublished formal language. *arXiv e-prints*, pages arXiv–2305.

Weber, F., Wambsganss, T., Rüttimann, D., and Söllner, M. (2021). Pedagogical agents for interactive learning: A taxonomy of conversational agents in education. In *Forty-Second International Conference on Information Systems. Austin, Texas*, pages 1–17.

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., et al. (2023). A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Yu, W., Jiang, Z., Dong, Y., and Feng, J. (2020). Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.