Acting Humanly: Identification and Analysis of Logical Reasoning Biases Exhibited by ChatGPT versus Undergraduate Students

Augusto C. A. Oliveira[1], **Francisco L. B. Martins**[1], **Davi R. Vasconcelos**[1], **Maria V. Menezes**[1]

[1]Universidade Federal do Ceará (UFC) - Campus Quixadá
Quixadá – CE – Brazil

{augustces,lmartins}@alu.ufc.br, {daviromero, vivianemenezes}@ufc.br

Abstract. *Definitions of Artificial Intelligence (AI) include characterizing algorithms as those that: thinking humanly, thinking rationally, acting humanly and acting rationally. On the one hand, Logic, as a formal framework, allows for the creation of algorithms capable of thinking rationally by expressing real world situations in a language that enables valid and rigorous reasoning. On the other hand, Large Language Models, such as ChatGPT, represent algorithms that acting humanly, especially in tasks involving understanding and generating natural language text. However, these models can exhibit logical reasoning biases, which are tendencies that impair the ability to reason logically. This article aims to identify and analyze the logical reasoning biases exhibited by ChatGPT in comparison to those exhibited by Information Technology Undergraduate Students, beginners in the Logic course.*

1. Introduction

Definitions of Artificial Intelligence (AI) can be categorized into algorithms that [Russell and Norvig 2016]: (i) thinking humanly; (ii) thinking rationally; (iii) acting humanly; and (iv) acting rationally. [Bellman 1978] considers *thinking humanly* means "automating activities that we associate with human thinking, such as problem-solving and learning". However, to claim that a program thinks like a human, it is necessary to understand how humans think [Russell and Norvig 2016]. According to [Winston 1992], an algorithm that *thinks rationally* "performs computations that make it possible to perceive, reason, and act". Logic [Whitehead and Russell 1927] is seen as a representation of this category, aiming to express world situations in formal language so that we can reason about them with rigorously defended arguments [Huth and Ryan 2004]. An AI that *acting humanly* [Rich and Knight 1991] "studies how algorithms can perform tasks that are currently better done by people". Recently, we have seen a massive adoption of Large Language Models (LLMs) [Lewis et al. 2020], which are considered representations of this category, particularly in natural language understanding and text generation tasks. Finally, algorithms *acting rationally* aim to achieve the best possible outcome.

However, despite the widespread use of LLMs in various applications, these models can provide incorrect responses to user inputs, especially on tasks involving logical reasoning [Ando et al. 2023, Liu et al. 2023, Martins et al. 2023]. According to [Han et al. 2024], there are studies in the literature aimed at evaluating how the capabilities of LLMs *compare to those of humans* in linguistic skills and others that aim at common sense knowledge and *logical reasoning* [Ando et al. 2023]. The work

[Gupta et al. 2024] aims to categorize the logical reasoning biases exhibited by LLMs. Logical reasoning biases are systematic deviations in how people interpret information and make decisions using mental heuristics. In other words, logical reasoning biases refer to mental tendencies that influence the ability to reason logically [Tversky et al. 1982].

This paper proposes to identify and analyze the logical reasoning biases exhibited by ChatGPT, a very popular LLM, in the task of reasoning in propositional logic, compared to the errors caused by Information Technology (IT) undergraduate students logical reasoning biases, who have not yet received formal instruction in logic, aiming to evaluate whether they naturally commit these biases. In order to achieve this, we will: (i) use a set of exercises answered by 247 undergraduate students; (ii) get the questions with the highest error rates and the possible logical reasoning biases involved in these errors; (iii) submit the same questions to ChatGPT; (iv) categorize the errors according to logical reasoning biases; and finally, (v) compare the biases done by the students with those exhibited by the tool. The remainder of the text is organized as follows: Section 2 presents the foundations; Section 3 shows related works; Section 4 presents the methodology; Section 5 shows the results and analysis of student performance versus the ChatGPT performance; and Section 6 presents the conclusions and future work.

2. Foundations

2.1. Propositional Logic

Propositional Logic is based on propositions, or declarative sentences, which one can argue as being true or false. It provides a solid foundation for the development of formal reasoning and the construction of computational systems. We consider certain declarative sentences as being *atomic* (propositional atom), like the sentence "*the dollar has risen*" .

Definition 1 *(Propositional Logic Language) The set of propositional atoms, together with the connectives (negation $\neg$, conjunction $\wedge$, disjunction $\vee$, conditional $\rightarrow$) and the symbols '(', ')', composes the alphabet of the propositional logic language, which can be defined by a grammar in Backus Naur form (BNF), as follows [Huth and Ryan 2004]:*

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

where P stands for any atomic proposition and each occurrence of φ to the right of $::=$ stands for any already constructed formula. We call $\mathcal{L}_{PL}$, named language of propositional logic, the set of all formulas obtained by BNF grammar.

For example, we can use the atom P to represent "*the dollar has risen*" and the propositional atom Q to represent "*the products become more expensive*". In order to encode more complex sentences, we use logical connectives. For example: $\neg P$ encodes the sentence "*the dollar has not risen*"; $P \wedge Q$ encodes the sentence "*the dollar has risen and the products have become more expensive*"; $P \vee Q$ encodes the sentence "*the dollar has risen or the products have become more expensive*" and; $P \rightarrow Q$ encodes the sentence "*if the dollar has risen then the products have become more expensive*".

The semantics of propositional logic assigns truth values to formulas, which can be either true (T) or false (F). The truth value of a formula depends on the truth values assigned to propositional atoms, as in Definition 2, and the meaning of logical connectives, as showed in Definition 3. For example, the truth value of the formula $P \wedge Q$ is given by the truth values of the atoms P and Q, along with the semantics of the conjunction $\wedge$.

Definition 2 *(Valuation for Atoms) Let $\mathbb{P}$ be a set of propositional atoms. The valuation for $P \in \mathbb{P}$ is defined as $v : \mathbb{P} \mapsto \{F, T\}$ where $v(P) = F$ or $v(P) = T$ [Enderton 2001].*

Definition 3 *(Valuation for Formulas) Let φ be a propositional logic formula and let $v : P \mapsto \{F, T\}$ be the valuation function for propositional atoms. The valuation function for propositional formulas is defined as $\bar{v} : \mathcal{L}_{LP} \mapsto \{F, T\}$ [Enderton 2001]:*

1. $\bar{v}(\mathcal{P}) = v(\mathcal{P})$.
2. $\bar{v}((\neg\varphi)) = \begin{cases} T, & \textit{if } \bar{v}(\varphi) = F \\ F, & \textit{otherwise.} \end{cases}$
3. $\bar{v}((\varphi \wedge \psi)) = \begin{cases} T, & \textit{if } \bar{v}(\varphi) = T \quad \textit{and} \quad \bar{v}(\psi) = T \\ F, & \textit{otherwise.} \end{cases}$
4. $\bar{v}((\varphi \vee \psi)) = \begin{cases} T, & \textit{if } \bar{v}(\varphi) = T \quad \textit{or} \quad \bar{v}(\psi) = T \\ F, & \textit{otherwise.} \end{cases}$
5. $\bar{v}((\varphi \to \psi)) = \begin{cases} T, & \textit{if } \bar{v}(\varphi) = F \quad \textit{or} \quad \bar{v}(\psi) = T \\ F, & \textit{otherwise.} \end{cases}$

A valid logical reasoning represents a semantic entailment, establishing this relationship if for all cases where a set of formulas $\varphi_1, \varphi_2 \cdots, \varphi_n$ has a true truth value, then the conclusion ψ also has a true truth value, as presented in Definition 4.

Definition 4 *(Semantic Entailment) If, for all valuations in which all $\varphi_1, \varphi_2 \cdots, \varphi_n$ evaluate to true, ψ evaluates to true as well, we say that $\varphi_1, \varphi_2 \cdots, \varphi_n \models \psi$ holds and call $\models$ the semantic entailment relation [Huth and Ryan 2004].*

In the example depicted in Figure 1 we want to know if the statments "*if the dollar has risen, then the products become more expensive*" and "*the products have not become more expensive*" semantically entails "*the dollar has risen*". The statement "*if the dollar has risen, then the products become more expensive*" can be encoded as $P \to Q$ and; "*the products did not become more expensive*", as $\neg Q$. We need to verify if $P \to Q, \neg Q \models \neg P$. Thus, we have to check if for all valuations that $P \to Q, \neg Q$ are true, $\neg P$ is also true. In the only one case where $\bar{v}(P \to Q) = T$ and also $\bar{v}(\neg Q) = T$, we have $v(P) = F$ and $v(Q) = F$. In this case, $\bar{v}(\neg P) = T$. Thus, we prove that $P \to Q, \neg Q \models \neg P$.

Figure 1. An example of propositional logic reasoning exercise.

Consider the following statements:
1. If the dollar has risen, then the products become more expensive;
2. The products did not become more expensive.
Can we conclude that the statement below logically follows from the statements above?
3. The dollar did not rise.

2.2. Biases in Logical Reasoning

A bias in logical reasoning [Tversky et al. 1982, Bennett 2012] is described as systematic deviations in how people interpret information and make decisions. These biases can lead to inaccurate judgments and incorrect predictions. In the book [Bennett 2012], the authors describe more than 300 logical reasoning biases. We highlight two of them: **affirming the**

consequent and **denying the antecedent**. Both biases refer to conditional formulas such as $\varphi \to \psi$, where the formula φ is called *antecedent* (the formula after $\to$) and formula ψ is called **consequent** (the formula before $\to$). Affirming the consequent bias is a logical reasoning error where in a conditional statement if the consequent is considered true, then the antecedent is also considered true. On the other hand, denying the antecedent bias is a logical reasoning error where in a conditional statement, if the antecedent is false, it is erroneously concluded that the consequent is also false. Considering the propositional atoms P, representing the statement "*the dollar has risen*", and Q, representing the statement "*the products become more expensive*", observe the following examples.

Example 1 *(Affirming the consequent bias) Consider the statements presented in Figure 2 and the formulas* $P \to Q$, Q, *representing the statements 1 e 2, and* P *representing* 3. *According to the this bias, one erroneously concludes that* P *is true ("the dollar has risen") given that we have* $P \to Q$ *and the affirmation of its consequent* Q *(*Q *is true). However, when* $v(P) = F$ *and* $v(Q) = T$*, then* $\bar{v}(P \to Q) = T$. *Thus, according to the Definition 4,* $P \to Q, Q \not\models P$. *So, we cannot logically conclude* P *from* $P \to Q$ *and* Q.

Figure 2. Logical reasoning exercise - Question Q3.

Consider the following statements:
1. If the dollar has risen, then the products become more expensive;
2. The products became more expensive.
Can we conclude that the statement below logically follows from the statements above?
3. The dollar rose.

Example 2 *(Denying the antecedent bias) Consider the statements presented in the Figure 3 and the formulas* $P \to Q$, $\neg P$, *representing statements 1 and 2, and* Q *representing the conclusion* 3. *According to this bias, one can incorrectly concludes that* $\bar{v}(\neg Q) = T$ *("the products did not became more expensive") given that* $\bar{v}(P \to Q) = T$ *and* $\bar{v}(\neg P) = T$. *However, when* $v(P) = F$ *and* $v(Q) = T$*, we have* $\bar{v}(P \to Q) = T$, $\bar{v}(\neg P) = T$ *and* $\bar{v}(\neg Q) = F$. *Thus, according to the Definition 4, we have* $P \to Q, \neg P \not\models Q$. *So, we cannot logically conclude* $\neg Q$ *from the statements* $P \to Q$ *and* $\neg P$.

Figure 3. Logical reasoning exercise - Question Q4.

Consider the following statements:
1. If dollar has risen, then products become more expensive;
2. The dollar did not rise.
Can we conclude that the statement below logically follows from the statements above?
3. Products did not become more expensive.

2.3. Large Language Models

Machine learning (ML) [Russell and Norvig 2016] is a subarea of AI that studies algorithms and models capable of learning from data. One of the most significant advances in ML has been the development of large language models (LLMs), such as OpenAI's GPT (Generative Pre-trained Transformer) [OpenAI 2021]. These models are trained on vast amounts of text, allowing them to capture a wide range of knowledge about natural language. In 2020, the 3rd generation GPT, GPT-3, was released [Zhang and Li 2021]. GPT-3.5, released in 2022, ensured advances in performance and safety, offering more accurate and coherent responses. Finally, GPT-4, released in 2023, introduced a rule-based reward model approach, along with reinforcement learning with human feedback, improving the model's performance and safety compared to its previous versions [Koubaa 2023].

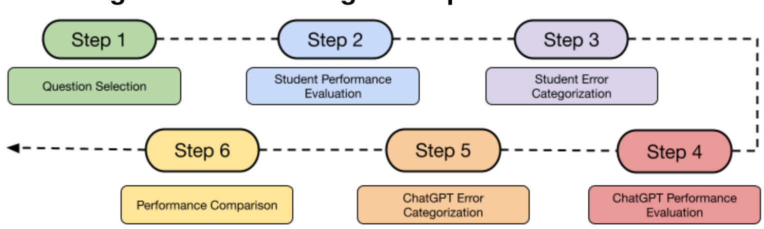
3. Related Work

In the work [Ando et al. 2023], the authors investigate if LLMs exhibit logical reasoning biases similar to human-like biases, focusing on reasoning about syllogisms in predicate logic. In order to evaluate these biases, the authors used a dataset [Shikishima et al. 2009], which compiles studies on human inference skills with syllogistic inference questions and manually assigned labels such as contradiction, implication, or neutral. RoBERTa [Liu et al. 2019], BART [Lewis et al. 2020] and GPT-3.5 models were evaluated, and the results indicated that they exhibit failures caused by biases similar to human ones. Secondly, the paper [Yang et al. 2024] explores the iterative theorem provers area, where human experts interact with tools (e.g., Coq [Coq 1996], Isabelle [Nipkow et al. 2002], Lean [De Moura et al. 2015]) for constructing mathematical proofs. The authors discuss how LLMs can be used to automate these interactions. They propose a tool for theorem proving that extracts data from *Lean* for premise annotations and creates a new theorem benchmark. It outperformed a GPT-4-based baseline and proved 51.2% of theorems, including 65 theorems without existing proofs in *Lean*. Finally, the work [Saparov et al. 2023] evaluates the deductive reasoning capabilities of LLMs, according their ability to learn from simple proofs and generalize this learning to more complex proofs. The evaluation was conducted with a dataset that uses fictitious names to prevent the exploitation of pre-trained knowledge. The models *FLAN-T5* [Chung et al. 2024], *LLaMA* [Touvron et al. 2023], PALM and *GPT-3.5* were evaluated. The results showed that the models utilized reasonably well natural deduction rules and were able to generate compositional proofs but struggled with longer proofs and hypothetical sub-proofs, especially in proofs by cases and contradiction.

4. Methodology

This work aims to evaluate whether ChatGPT exhibits logical reasoning biases similar to those performed by undergraduate students. To do this, we will follow the methodological steps depicted in Figure 4 and described as follows.

Figure 4. Methodological steps of the research.



Step 1: Question Selection. We used a set of 12 reasoning exercises in propositional logic. The exercises contain a set of statements and a conclusion. In each exercise, it is asked whether a conclusion logically follows from a set of statements. Figure 1, 2 and 3 illustrate examples of logical reasoning exercises. Students must answer "*yes*" if they believe the conclusion logically follows from the statements, or "*no*" if they believe

it does not. These exercises were given on the first day of the Logic in Computer Science course, offered to IT students in the semesters 2023.1, 2023.2, and 2024.1. The exercises and their correct answers are available in `https://x.gd/7bKDX`. **Step 2: Students Performance Evaluation**. We evaluated the performance of 247 students who answered the exercises. In this evaluation, the goal is to identify the questions with the highest error rates. **Step 3: Students Errors Categorization**. We categorized, according to logical reasoning biases, the errors performed by students in the questions with the highest error rates. **Step 4: ChatGPT Performance Evaluation**. We asked ChatGPT to answer the propositional logic reasoning exercises. We used versions 3.5 and 4.0 of the tool. Interactions with the bot were conducted through the prompt interface, where we submitted each question and saved the answers. Finally, we evaluated the answers to determine which questions were answered correctly and which ones the bot answered incorrectly. **Step 5: ChatGPT Errors Categorization**. We categorized the logical reasoning errors exhibited by the tool according to logical reasoning biases. **Step 6: Performance Comparison**. We conducted two comparisons: (i) in the first, we compare the performance of the evaluated bot versions in terms of accuracy rate and quality of the responses provided; (ii) in the second, we compare the performance of students regarding to logical reasoning biases and the ChatGPT performance related to these biases.

5. Results and Analyses

5.1. Students Performance

After evaluating the exercises of the 247 students, we identified the questions with the highest number of errors as follows: Q3 (Figure 2) with 63.16$\%$ incorrect answers; Q4 (Figure 3 with 49.39$\%$ incorrect answers; **Q9** (Figure 5) with **65.59**$\%$ incorrect answers and; **Q12** (Figure 6) with **61.54**$\%$ incorrect answers. The complete evaluation with all correct and incorrect answers for all questions and the total number of students is available at `https://x.gd/TpgEs`. Next, we categorized the logical reasoning errors for each question based on logical reasoning biases, as illustrated below:

- **Q3: affirming the consequent bias** Figure 2 depicts Question Q3. The errors were categorized according to affirming the consequent bias. According to this bias, if the consequent of the conditional "*the products become more expensive*" is true, there is a tendency to believe it is logically correct to affirm that the antecedent of the conditional is also true ("*the dollar has risen*"). The proof that statements $1, 2$ do not semantically entail 3 is presented in Example 1.
- **Q4: denying the antecedent bias** Figure 3 shows Question Q4. The errors exhibited in this question were categorized according to the denying the antecedent bias. According to this bias, if the antecedent of the conditional statement "*the dollar has risen*" is false ("*the dollar did not rise*"), there is a tendency to believe it is logically correct to affirm that the consequent of the conditional statement is also false ("*the products did not become more expensive*"). The proof that statements $1, 2$ do not semantically entail 3 is presented in Example 2.
- **Q9: denying the antecedent bias** Figure 5 shows Question Q9. The errors exhibited in this question were categorized according to the denying the antecedent bias. According to this bias, if the antecedent of the conditional statement "*the dollar has risen or the oil has risen*" is false ("*the dollar did not rise*" and "*the*

oil did not rise"), there is a tendency to believe it is logically correct to affirm that the consequent of the conditional statement is also false ("*the inflation did not increase*"). In this question, the representation of the statements in propositional atoms can be given as follows: P: "*the dollar has risen*"; Q: "*the oil has risen*" and R: "the inflation increases". Thus, we need to verify if $\neg R$ logically follows from $P \vee Q \to R, \neg P, \neg Q \models \neg R$. Notice that, when $v(P) = F$, $v(Q) = F$ and $v(R) = T$, we have $\bar{v}((P \vee Q) \to R) = T$, $\bar{v}(\neg P) = T$ and $\bar{v}(\neg Q) = T$, while $\bar{v}(\neg R) = F$. Therefore, by Definition 4, $(P \vee Q) \to R, \neg P, \neg Q \not\models \neg R$. Hence, the statement "*inflation did not increase*" does not logically follow from 1, 2, 3.

Figure 5. Logical reasoning exercise - Question Q9.

Consider the following statements:
1. If the dollar has risen or the oil has risen, then the inflation increases;
2. The dollar did not rise;
3. The oil did not rise.
Can we conclude that the statement below logically follows from the statements above?
4. The inflation did not increase.

- **Q12: affirming the consequent and denying the antecedent biases**. Figure 6 shows Question Q12. The errors made in this question were categorized according to affirming the consequent and denying the antecedent biases. According to the affirming the consequent bias, the fact that the consequent of the first conditional "*the products became more expensive*" is true leads to the tendency to believe that it is logically correct to affirm that the antecedent of the first conditional "*the dollar has risen*" is also true. Conversely, assuming that the antecedent of the first conditional "*the dollar has risen*" is true makes the antecedent of the second conditional "*the dollar did not rise*" false. According to denying the antecedent bias, concluding that the antecedent of the second conditional "*the dollar did not rise*" is false tends to lead to the belief that it is logically correct to affirm that the consequent of the second conditional is also false ("*I do not buy more food*"). In this question, the representation of the statements in propositional atoms can be given as follows: P: "*the dollar has risen*", Q: "*the products became more expensive*", R: "*I buy more food*". Thus, we need to verify if $P \to Q, \neg P \to R, Q \models \neg R$. Notice that when $v(P) = F$, $v(Q) = T$ and $v(R) = T$, then $\bar{v}(P \to Q) = T$, $\bar{v}(\neg P \to R) = T$ and $v(Q) = T$, while $\bar{v}(\neg R) = F$. Therefore, by Definition 4, $P \to Q$, $\neg P \to R$, $Q \not\models \neg R$. Hence, "*I do not buy food*" does not logically follow from the statements 1, 2, 3.

Figure 6. Logical reasoning exercise – Question Q12.

Consider the following statements:
1. If the dollar has risen, then the products become more expensive;
2. If the dollar does not rise, then I buy more food;
3. The products became more expensive.
Can we conclude that the statement below logically follows from the statements above?
4. I did not buy more food.

5.2. ChatGPT Performance

We evaluated 12 exercises using the GPT-3.5 and GPT-4 versions. In GPT-3.5, the bot answered incorrectly in 5 out of the 12 exercises. Among the incorrect responses were those to questions Q3, Q4, Q9, and Q12.

- **Q3: affirming the consequent bias**. **Answer**: "*Yes, we can conclude that the dollar rose, because if the dollar rises, products become more expensive, and the products did become more expensive. Therefore, the statement that the dollar rose logically follows from the given statements.*"
  - **Answer analyses**. Given: P - "*the dollar has risen*" and Q - "*products became more expensive*". According to this answer, the bot considered $v(P) = T$ ("*we can conclude that the dollar rose*") and $v(Q) = T$ ("*and the products became more expensive*"). However, as shown in Section 5.1.1, this does not prove that $P \rightarrow Q, Q$ semantically entails P, being an example of the affirming the consequent bias.
- **Q4: denying the antecedent bias**. **Answer**. "*Yes, we can conclude that the products did not become more expensive, because the first conditional statement was not satisfied (the dollar did not rise), which prevents the consequence that the products would become more expensive".*
  - **Answer analyses**. Given: P - "*the dollar has risen*" and Q - "*the products became more expensive*". According to this answer, the bot considered $v(Q) = F$ ("*we can conclude that the products did not become more expensive*") by the fact that $v(P) = F$ ("*the first conditional statement was not satisfied.*"). However, despite these valuations making $\bar{v}(P \rightarrow Q) = T$, $\bar{v}(\neg P) = T$, they do not prove that $P \rightarrow Q$, $\neg P$ semantically entails $\neg Q$ (as show in Section 5.1.3). Thus, this is an example of an error due to denying the antecedent bias.
- **Q9: denying the antecedent bias Answer**. *Yes, we can conclude that inflation did not increase, as both necessary conditions for inflation to rise (the dollar rising and oil rising) were not met.*
  - **Answer analyses.** Given: P - "*the dollar rises*"; Q ''the oil rises'' and; R - "increase the inflation". *Bot* considers that $v(R) = F$ ("*we can conclude that inflation did not increase.*"), $v(P) = F$ and $v(Q) = F$ ("*dollar rising and oil rising were not met*"). However, despite these valuations making the formulas $\bar{v}((P \vee Q) \rightarrow \neg R) = T$, $\bar{v}(\neg P) = T$, $\bar{v}(\neg Q) = T$, $\bar{v}(\neg R) = T$, they do not prove that $(P \vee Q) \rightarrow \neg R$, $\neg P$, $\neg Q$ semantically entail $\neg R$ (as show in Section 5.1.3). Thus, we conclude that the bot made an error according to the denying the antecedent bias.
- **Q12: affirming the consequent and denying the antecedent biases**. **Answer**. "*Yes, we can conclude that the statement "I did not buy more food" logically follows from the given statements. This is because if products became more expensive due to the dollar rising, then a person might have chosen not to buy more food, even if the dollar did not rise".*
  - **Answer analyses.**. Given: P - "*the dollar has risen*"; Q - "*the products became more expensive*" and; R - "*I bought more food*". According to this answer, bot considered that $v(R) = F$ ("*statement "I did not buy more food" logically follows from the given statement*"); $v(Q) = T$ and

$v(P) = T$ ("*This is because if the products became more expensive due to the dollar rising...*"). Thus, bot considers that $\bar{v}(P \to Q) = T$, $\bar{v}(\neg P \to R) = T$, $\bar{v}(Q) = T$ and $\bar{v}(\neg R) = T$. However, this does not prove that $P \to Q, \neg P \to R \models Q$ (as showed in the Section 5.1.4). Stating that "*products became more expensive due to the dollar rising*" is an example of affirming the consequent bias, and stating that "a person might have chosen not to buy more food even if the dollar did not rise" is an error according to the denying the antecedent bias.

In the **GPT-4**, questions with incorrect answers were: Q9 and Q12. The responses and analyses are reported as follow.

- **Q9: denying the antecedent bias**. **Answer**. Yes, we can conclude that "inflation did not increase" from the given statements. Let's analyze the reasoning in propositional logic. Let's define: P - The dollar has risen;Q - The oil has risen; R - inflation increases. The statements can be expressed as: 1. $(P \vee Q) \to R$ (If the dollar has risen or oil has risen, then inflation increases.) 2. $\neg P$ (The dollar did not rise.) 3. $\neg Q$ (Oil prices did not rise.) We want to conclude if $\neg R$ logically follows from these statements. **Let's analyze the implication** $(P \vee Q) \to R$**. It tells us that if** $P \vee Q$ **is true, then** R **must be true**. Now, let's analyze $P \vee Q$: it means that at least one of the propositions P or Q is true. With $\neg P$ and $\neg Q$, we can conclude that $P \vee Q$ is false because P and Q are false. Given that $P \vee Q$ is false, the implication $(P \vee Q) \to R$ is true regardless of the value of R. This means that the antecedent $P \vee Q$ is false, which makes the implication $(P \vee Q) \to R$ true. However, **knowing that** $P \vee Q$ **is false, we see that the necessary condition for** R **has not been satisfied. Therefore,** R **cannot be true.** In other words, if P and Q are false, then inflation does not increase ($\neg R$). Therefore, the conclusion "inflation did not increase" logically follows from the statements.
  - **Answer analyses**. We highlighted in the response the texts where the *bot* makes errors related to the denying the antecedent bias: (i) "*the implication* $(P \vee Q) \to R$ *tells us if* $P \vee Q$ *is true, then* R *must be true*" and (ii) "*knowing* $P \vee Q$ *is false, we see that the condition for* R *has not been satisfied. Therefore,* R *cannot be true*". In both sentences, *bot* considers if the antecedent of the implication is false, the consequent must be false.
- **Q12: affirming the consequent and denying the antecedent biases**. **Answer**. Yes, we can conclude that "I didn't buy more food" from the given statements. Let's analyze the reasoning in propositional logic. First, let's define the propositions: P - The dollar rises; Q - Products become more expensive; R - I buy more food. The statements can be expressed as: *1.* $P \to Q$ *(If the dollar rises, then products become more expensive.) 2.* $\neg P \to R$ *(If the dollar does not rise, then I buy more food.) 3.* Q *(Products became more expensive.)* We want to conclude if $\neg R$ follows from these statements. Let's analyze: $P \to Q$ (If the dollar goes up, then products become more expensive.), Q (Products became more expensive.) From statement 3 (Q), we know that products became more expensive. **The implication** $P \to Q$ **tells us that if** P **is true, then** Q **must be true**. Since Q is true, this suggests that P might be true. Now, let's consider the contradiction with $\neg P \to R$: If $\neg P$, then R. If Q is true, this must imply that P is true, given that $P \to Q$ is a valid implication. Therefore, if Q is true, the only way consistent with

the statements is that P is true. If P is true, then $\neg P$ is false. If $\neg P$ is false, then R cannot be inferred because the necessary condition ($\neg P$) for R to be true is not satisfied. Therefore, we can conclude that $\neg R$ must be true. So, the conclusion "I didn't buy more food" ($\neg R$) logically follows from the statements.

- **Answer Analyses**. We have highlighted the sentences where logical reasoning biases are committed. In the sentence "The implication $P \rightarrow Q$ tells us that if P is true, then Q must be true. Since Q is true, this suggests that P might be true, "the **affirming the consequent** bias is committed, and inmust the sentence "*If P is true, then $\neg P$ (The dollar did not go up) is false. If $\neg P$ is false, then R (I buy more food) cannot be inferred because the necessary condition ($\neg P$) for R to be true is not satisfied. Therefore, we can conclude that $\neg R$ (I didn't buy more food) must be true*" the error due to the **denying the antecedent** bias is committed. In the first sentence, the *bot* considers that, since the consequent is true, the antecedent must also be true. In the second sentence, the *bot* considers that, since the antecedent is false, the consequent must also be false.

5.3. Performance Comparison

5.3.1. *Bot* versus *Bot*

In this section, we briefly compare the answers given by ChatGPT in its versions 3.5 and 4.0 of. In version 3.5, the *bot* got 7 out of 12 exercises correct (58.33% accuracy), while in version 4.0, the *bot* got 10 out of 12 exercises correct (83.33% accuracy). Besides improving the accuracy rate, the bot provided more comprehensive answers, formalizing sentences into propositional atoms and then into formulas with connectives. We observe this improvement also in a properly supported justifications from a logical point of view. However, as shown in the previous section, even with an improvement in formalization, the bot still commits errors related to logical reasoning biases in its justifications.

5.3.2. *Human* versus *Bot*

Comparing the undergraduate students perfomance versus *bot* performance in the task of solving the proposed problems, we conclude: (i) questions with the highest error rates among students (Q3, Q4, Q9, and Q12) were also answered incorrectly by GPT-3.5, and questions Q9 and Q12 were similarly answered incorrectly by GPT-4; (ii) justifications given by the *bot* in its version 3.5 (in questions Q3, Q4, Q9 and Q12) were relate to the logical reasoning biases: affirming the consequent and denying the antecedent. Such biases were possibly committed by the students who answered these questions incorrectly and; (iii) justifications given by the *bot* in its version 4.0, in questions Q9 and Q12, were relate to the same logical reasoning biases: affirming the consequent and denying the antecedent. These biases were committed by the *bot* in the provided answers, despite a better presentation of the justifications from a logical point of view.

6. Conclusions and Future Work

This paper presented a comparison between the performance of undergraduate students and ChatGPT (GPT-3.5 and GPT-4) in solving 12 exercises on propositional logic reasoning named Q1, Q2, $\cdots$ Q12. Results showed that both humans and the AI model made similar errors related to logical reasoning biases: affirming the consequent and denying the antecedent. A total of 247 students answered the exercises and highest error questions rates were Q3 (63.16%), Q4 (49.39%), Q9 (65.59%) and Q12 (61.54%). These questions were associated with the mentioned biases, highlighting a common difficulty among students in avoiding such logical reasoning errors. ChatGPT showed an evolution between the GPT-3.5 and GPT-4.0 versions in terms of the number of correct answers. In GPT-3.5 version, the model got 58.33% of the questions correct, while GPT-4 version achieved an accuracy rate of 83.33%. However, questions Q9 and Q12 were missed by both versions of the model, suggesting that although the model has improved, it still exhibits logical reasoning biases similar to humans. As future work, we intend to test the model performance in solving exercises involving another deductive systems such as analytic tableaux.

References

Ando, R., Morishita, T., Abe, H., Mineshima, K., and Okada, M. (2023). Evaluating large language models with NeuBAROCO: Syllogistic reasoning ability and human-like biases. In Chatzikyriakidis, S. and de Paiva, V., editors, *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pages 1–11, Nancy, France. Association for Computational Linguistics.

Bellman, R. (1978). *An Introduction to Artificial Intelligence: Can Computers Think?* Boyd & Fraser Publishing Company.

Bennett, B. (2012). *Logically fallacious: the ultimate collection of over 300 logical fallacies (Academic Edition)*. eBookIt. com.

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

Coq, P. (1996). The coq proof assistant-reference manual. *INRIA Rocquencourt and ENS Lyon, version*, 5.

De Moura, L., Kong, S., Avigad, J., Van Doorn, F., and von Raumer, J. (2015). The lean theorem prover (system description). In *Automated Deduction-CADE-25: 25th International Conference on Automated Deduction, Berlin, Germany, August 1-7, 2015, Proceedings 25*, pages 378–388. Springer.

Enderton, H. B. (2001). *A mathematical introduction to logic*. Elsevier.

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., and Khot, T. (2024). Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Han, S. J., Ransom, K. J., Perfors, A., and Kemp, C. (2024). Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83:101155.

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

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

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Martins, F., Oliveira, A., Vasconcelos, D., and Menezes, M. (2023). Avaliando a habilidade do chatgpt de realizar provas de dedução natural em lógica proposicional. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1282–1292, Porto Alegre, RS, Brasil. SBC.

Nipkow, T., Wenzel, M., and Paulson, L. C. (2002). *Isabelle/HOL: a proof assistant for higher-order logic*. Springer.

OpenAI (2021). ChatGPT. `https://openai.com/research/chatgpt`. Accessed on: April 3, 2024.

Rich, E. and Knight, K. (1991). *Artificial Intelligence*. Artificial Intelligence Series. McGraw-Hill.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

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. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 3083–3105. Curran Associates, Inc.

Shikishima, C., Hiraishi, K., Yamagata, S., Sugimoto, Y., Takemura, R., Ozaki, K., Okada, M., Toda, T., and Ando, J. (2009). Is g an entity? a japanese twin study using syllogisms and intelligence tests. *Intelligence*, 37(3):256–267.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.

Tversky, A., Kahneman, D., and Slovic, P. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge.

Whitehead, A. and Russell, B. (1927). *Principia Mathematica*. Number v. 1 in Cambridge mathematical library. Cambridge University Press.

Winston, P. H. (1992). *Artificial intelligence (3rd ed.)*. Addison-Wesley Longman Publishing Co., Inc., USA.

Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., and Anandkumar, A. (2024). Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, 36.

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