Experiment using ChatGPT to detect and analyze fallacious arguments

The following is a step-by-step we followed to prompt ChatGPT (both 3.5 and 4.0 versions).

Step 1: We opened a new chat and directly copied the prompt in table 1 below.

Estrutura do Prompt 01:
Extract all the Fallacious Argument Graph Units that this text contains and classify each fallacy. Indicate how to correct or avoid the fallacy when the author rewrites the AGU. • A fallacious AGU is defined as: Fallacious Argument Unit (AU) and unique handles. • An Argument Unit needs to have: one Claim (main point), and possibly one or more of the following elements: Data (evidence or grounds supporting the claim), Warrant (reasoning that connects the data to the claim), Backing (additional support for the warrant) or Rebuttal (addressing potential counterarguments). The AU is synthesized from the input text. • A handle is defined as: string quote from the input text, which can be used to find and verify the text from which the AU was synthesized. Building the Argument graph: • Extract the fallacious AGUs even if the conversation is informal. • Classify the type of fallacy found in each fallacious AU extracted and explain why it´s a fallacious argument unit. The Argument graph will be structured as follows: START • Fallacious argument graph unit (AGU) - synthetized argument:…, type of fallacy:…, explanation:… (handle: "…), fallacy correction:… END The text to be analyzed will be provided next.

Table 1. Structure of the prompt used in the experiment to detect, extract, classify, explain the fallacy, quote and suggest correcting the fallacious argument.

We were able to conclude that the text of the prompt in English worked best, and only the last sentence "The text to be analyzed will be provided below" in Portuguese resulted in a structured response.

Step 2: We copied a part of the text that had already been divided into eight, containing less than 4,000 characters without spaces.

ChatGPT already provided a first response, as in the following example:

ChatGPT 3.5 Answer
START • Fallacious argument graph unit (AGU) **Synthesized argument**: O nosso congresso foi muito maior do que o do partido socialista, portanto somos melhores. **Type of fallacy**: Relative privation (also known as the "not as bad as" fallacy). **Explanation**: This argument attempts to downplay the significance of a problem by comparing it to a worse problem. Just because one congress was larger than another does not necessarily mean the party is better or more successful. (**Handle**: "os dois maiores partidos portugueses tiveram um congresso em 2024, o nosso foi muito maior do que o do partido socialista."),

correção da falácia: The size of a congress does not determine the success or quality of a political party. A more valid comparison would involve the achievements and policies of each party.

Synthesized argument: Tornamo-nos na terceira maior força política deste país, a mais audível e a mais dinâmica, portanto merecemos vencer as eleições legislativas. **Type of fallacy**: Appeal to popularity (Argumentum ad populum). **Explanation**: This fallacy asserts that a claim is true or valid because a large number of people believe it to be so. However, popularity does not necessarily equate to correctness or suitability for governing. (**Handle**: "Tornamo-nos na terceira maior força política deste país, a mais audível e a mais dinâmica."), correção da falácia: Winning elections should be based on policies, qualifications, and track record rather than solely on popularity or size.

Synthesized argument: Só acerta em tudo quem nunca vai a jogo, portanto, porque fomos à luta, somos melhores do que os que ficaram no sofá. **Type of fallacy**: False dilemma (also known as black-or-white fallacy). **Explanation**: This fallacy presents a situation as having only two alternatives, when in fact, there are more. It oversimplifies a complex issue. (**Handle**: "Só acerta em tudo quem nunca vai a jogo, só acerta em tudo quem nunca vai à luta, e nós fomos."), correção da falácia: There are many ways to contribute to a cause, and not all of them involve direct action or participation in a specific activity like political campaigning. The implication that those who don't actively engage in the same way are inferior is flawed. END.

Table 2. Example of the first response given by ChatGPT 3.5 to the prompt.

Step 3: We used a second prompt to check whether ChatGPT could extract more fallacies.

The following prompt was applied until a repeated response appeared. This was necessary due to the size of the text inputed in ChatGPT. Since your dialog box has a character limitation, it is interesting to prompt ChatGPT again to obtain more responses if possible.

Structure of Prompt 02:
Can you identify more Argument Graph Units (AGUs) that may contain a fallacy in the text provided and that have not yet been analyzed? Do not identify or analyze arguments that you have already identified and classified in the previous question.

Table 3. Structure of the prompt to check for more fallacious arguments.

After 3 or 4 responses with different fallacious arguments detected, ChatGPT begins to show repetitions.

Step 4: We used a third prompt to confirm that it wouldn't find any more fallacious arguments.

Structure of Prompt 03:
You've already classified this unit. Is there anything else you could extract from the text at the beginning of this chat? If there isn't, say that there isn't.

We tested this prompt on ChatGPT-3.5 directly from the OpenAi[1] website and on ChatGPT-4 (OpenAI, 2023), using Co-Pilot tool (Microsoft Corporation, 2021), available on Microsoft's search engine, Bing.

Natural language prompts:

Repetimos o experimento com uma diferença: utilizando prompts com linguagem natural, como em uma conversação, mas pedindo para o GPT-4 encontrar falácias e explicá-las. Não pedimos para que fornecesse uma citação do texto e uma correção do argumento para não sobrecarregar as instruções e confundir as respostas.

Prompt 01:
Is it possible to find fallacious arguments in the following text? Explain which fallacy and why or why not. Text:
Prompt 02 (aplicado a cada resposta, até que aparecesse uma resposta repetida).
Is it possible to identify more Argument Units that may contain a fallacy in the text provided and that have not yet been analyzed? Do not identify or analyze arguments that you have already identified and classified in the previous question.
Prompt 03 (aplicado após uma resposta contendo pelo menos uma argumente unit repetida.
You've already classified this unit of argument. Are there any others that you could extract from the text at the beginning of this chat? If there isn't, tell me there isn't.

---

[1] OpenAi, Chat GPT 3.5: https://chat.openai.com/