The Fallacy of Explainable Generative AI: evidence from argumentative prompting in two domains

*Elena Musi[†], Rudi Palmieri[†]

Department of Communication and Media (University of Liverpool), School of the Arts, 19 Abercromby Square, L69 7ZG

Abstract

This contribution presents a methodology to investigate the soundness of GPT-4 explanations through a combination of fallacy theory and linguistic refinement. It seeks to address the following research questions: Can we leverage Argumentation Theory to i) elicit differences between LLMs' and human reasoning? ii) build prompting strategies to reduce hallucinations in explanations? To achieve this, we test four prompting strategies using GPT-4 across two domains (HR,loan), prompting the system to generate 30 explanations using analogical, causal, and counterfactual reasoning. We manually annotate the results to assess whether the justifications and the associated reasoning (argument scheme) are sound, fallacious and or influenced by contextual factors. Furthermore, we develop guidelines for prompt engineering to improve the argumentative quality of explanations.

Keywords

fallacy, GPT4, prompting strategies, explainability

1. Introduction

Even though news headlines are warning us everyday about the pitfalls of Generative Artificial Intelligence (e.g. hallucinations, bias etc), we are relying more and more on Generative AI to support strategic decision-making choices across different domains. In March 2024, the National Audit Office (NAO) reported that 70% of surveyed government bodies are either piloting or planning to use AI, with applications ranging from supporting operational decision-making to enhancing internal processes ($https://www.nao.org.uk/wp-content/uploads/2024/03/use-of-artificial-intelligence-in-government.pdf$). As explained in the Alan Turing Institute report "AI and Strategic Decision-Making Communicating trust and uncertainty in AI-enriched intelligence" ($https://tinyurl.com/ec2jr3fr$), the use of AI amplifies the perception of uncertainties in decision making processes: the opaque nature of Generative Artificial Intelligence systems makes it difficult to understand how conclusions have been reached. From an argumentative perspective, daily decision-making processes (such as hiring someone, giving a loan, buying a product) become argumentative issues when their outcomes directly impact citizens. The outcome is a standpoint while the reasons supporting the decision-making are arguments. The reasoning process which allows the arguments to support the standpoint (argument scheme) can be more or less fallacious. When Generative AI (GAI) is involved in the decision-making, the second rule of a critical discussion [1] is violated by evading the burden of proof. Asking the system to provide arguments for its choices does not constitute a solution since Generative AI systems are "stochastic parrots" who do not mean what they say [2]. In other words, the nature of Generative AI systems as stochastic probabilistic models results in an epistemology different from that of human arguers: our inferential processes are driven by motives beyond probabilistic calculations. Scholarly efforts have focused on elaborating prompting strategies such as Chain of Thought [3] and Tree of Thoughts [4] to improve Large Language Models' performance on logical tasks mimicking human reasoning. However, as acknowledged by the authors themselves,

CMNA workshop '24, 17th September 2024

*Corresponding author.

elena.musi@liverpool.ac.uk (*. Musi); palmieri@liverpool.ac.uk (R. Palmieri)

https://www.liverpool.ac.uk/communication-and-media/staff/elena-musi/ (*. Musi);

https://www.liverpool.ac.uk/communication-and-media/staff/rudi-palmieri/ (R. Palmieri)

0000-0002-0877-7063 (*. Musi); 0000-0002-5122-3058 (R. Palmieri)

the emulation of the thought processes of human reasoners does not help answer whether the neural network is actually reasoning. Furthermore, the settings where these prompting strategies have been tested (e.g. arithmetic tasks) are different from the contexts where we exercise our informal logic capabilities. On these grounds, we aim at answering the following research questions: Can we leverage Argumentation Theory to i) elicit differences between LLMs' and human reasoning? ii) build prompting strategies to reduce hallucinations in explanations? To answer these research questions we conduct a pilot study, taking two decision-making domains as a case study: hiring for a tech startup and granting a loan for a commercial bank. We devise a suite of four types of prompting strategies to test the following hypotheses:

- H1 : The specific domain of decision-making influences the choice of the argument scheme based on which the decision is justified.
- H2 : Providing domain-related information in the prompt influences the preference for a type of argument scheme.
- H3 : The justification advanced by the LLM is not always based on the purported argument scheme.
- H4 : The justification advanced by the LLM contains the use of fallacious arguments
- H5 : Prompts embedding critical questions in a tree of thought ('critically thought') lead to a lower number of hallucinated argument schemes and/or fallacious justifications than"un-critically thought" prompts.
- H6 : Prompts embedding critical questions ('critically thought') in both a tree of thought and chain of thought environment lead to a lower number of hallucinated argument schemes and/or fallacious justifications than prompts without chain of thought.

.

The paper is organized as follows: in section 2 we summarize state of art research about Argumentation and Large Language Models. We then describe the data and methods used in the pilot study introducing the notion of "critically thought" prompting. We then present the results of the analysis. Finally, the implications vis-à-vis hypotheses and research questions are discussed.

2. Related Work

Although the use of Argumentation theories and methods to achieve explainable Artificial Intelligence has a long tradition [5], the application of Argumentation to explain decisions taken by Large Language Models is still at its infancy. The scholarly community has so far focused on the analysis of the outputs of LLMs, on how to improve them and on their use to carry out classification tasks. To cite a few, Herbold et al. [6] carry out a a large-scale study comparing human-written versus ChatGPT-generated argumentative student essays showing that the latter are rated as higher in quality. Other studies focus on the optimization of LLMs to complete arguments [7]. From a qualitative perspective, Hinton and Wagemans [8] argue that LLMs outputs are less persuasive than human ones highlighting a set of weaknesses. When it comes to classification capabilities, Ruiz-Dolz and Lawrence [9] show that LLMs do not achieve high performance, for fallacy identification, questioning their reliability in the Wild. Flipping the coin from classification to production of reasoning Du et al. [10] propose a "society of minds" approach where multiple language models discuss with each other claims and reasoning processes over multiple rounds to arrive at a common answer, showing improved performance. Besides being computationally costly, this approach has been verified on a set of mathematical tasks that are far away from our daily decision making processes. To our knowledge Freedman et al. [11] have been pioneers in investigating the relation between arguments and explainability in LLMS for decision making tasks. They propose a method through LLMs to construct argumentation frameworks where arguments pro and con are considered and assigned argumentative strength scores, serving as the basis for formal reasoning in decision-making.

3. Pilot study design

For our pilot study we have focused on two scenarios: hiring for a tech company and granting a loan for a commercial bank. The system roles and tasks to be performed were, respectively: HR for a tech company – choose a candidate for a data scientist position and Personal loan officer – choose an applicant to whom to offer a loan. The analytic steps that we followed are the following: we elaborated four different types of prompts looking at domain features to induce decision making choices patterned with explanations; we then run the prompts for 30 iterations each using the OpenAI API GPT4 model (default parameters). We manually annotated the explanations at different levels to identify patterns. Finally, we carried out an inter-level and inter-domain comparison of the attested patterns.

3.1. Argumentative Prompting Strategies

We have devised 4 argumentative prompts shared across the HR and the loan domain designed to test the hypotheses. All the prompts are zero-shot since they do not contain examples of the task to be performed. All the prompts ask to justify the choice in a tailored manner: "Your justification must be based on one of the following types of reasoning: analogy, cause-effect, counterfactual. Your justification is three sentences long". We kept the parameter of temperature at 1 to avoid both redundancy and excessive randomness. The types of reasoning correspond to the argumentative notion of argument schemes [12] the inference linking arguments (justification) to the standpoint (chosen applicant). In the informal logic tradition, there is a proliferation of types of argument schemes. However, from a semantic-ontological perspective, argument schemes can be divided into three major classes [13]. We have chosen these three argument schemes as representative of the three classes:

- Intrinsic argument schemes: the ontological relation between the content of the premise and the content of the conclusion belong to the same semantic frame – example: cause-effect
- the ontological relation between the content of the premise and the content of the conclusion belong to different semantic frames – the existence of one state of affairs is independent from the existence of the other – example: analogy
- the reasoning includes elements of both intrinsic and extrinsic relations – example counterfactual (cause+alternative)

The first prompt 1 is the only one to include data about the candidates available as evidence to the system. The data for both domains are fictional and have been elaborated by the authors on the basis of their knowledge and taking inspiration from existing data (e.g. Job adverts for data scientist positions published on the Adzuna job listing website). To prevent skewed results due to different amounts of training data associated with names and surnames we have chosen the most frequent male and female names in American English . We have intentionally kept a gender difference. The profile of the applicant was designed to avoid one applicant being an obvious choice with respect to the other.

The other prompts do not include data about the applicants since we want to test, in comparison with prompt 1, whether the domain of decision-making, regardless of the data provided, influences the choice of the argument scheme (H2). Additionally, Prompt 2 is an instance of Zero-Shot Chain of Thought Prompting [14]: the sentence "Let's think step by step" is added to induce the system eliciting the intermediate reasoning steps taken to justify the choice. In this way, we aim at observing whether CoT impacts the number of hallucinated and/or fallacious justifications (H3 and H4).

Prompt 3 takes as the basis Prompt 2 and it adds on critical questions [15] for each argument scheme to make the system evaluate whether the purported justification is fallacious or not. The critical questions for the causal and the analogical argument scheme draw from Tindale [16]. We call this prompting "critically thought" since we aim at testing whether the system is able to identify fallacies and whether the task of fallacies identification leads to less hallucinated and fallacious justifications (H5).

Prompt 4 differs from Prompt 3 since it is an instance of Tree of Thoughts Prompt [17, 4] – the system is prompted to explore a solution space with different alternatives with the possibility to backtrack during the decision-making process when one or more alternatives are evaluated unsuitable. In our case,

Textual Prompt HR	Textual Prompt Loan
Role:“You are the HR for a tech company. You need to choose a new Senior Data Scientist. The job description is the following: the Senior Data Scientist will work closely with a team of Data Engineers and Program Managers to solve real-world problems through state-of-the-art approaches using text, images, and other types of data. The professional qualities required are the following: experience with data scripting languages; Experience working as a Data Scientist; Experience with knowledge engineering; Focus on Natural Language processing (NLP), Machine Learning and Semantic Web/Ontology/Knowledge graph; A degree in Data Science, NLP or a STEM subject. You have received two applications". Content:“The first application is from James Smith. Current position: Data scientist. Degree: Computational Biology. Coding languages: Java and Python. Experience: building metrics for social media interactions. The Second application is by Mary Jones. Current position: Data scientist; Degree: Computational Social Science ; Coding languages: Python; Experience: building conversational agents. You choose to hire one of the two. You need to justify your choice. Your justification must be based on one of the following types of reasoning: analogy, cause-effect, counterfactual. Your justification is three sentences long".	Role: “You work in the personal loan office of a commercial bank. You have received two applications for a 15,000£ loan. Loan length: 5 years. APR: 8 %"; Content: " Application 1 is from James Smith. Reason for loan: home improvement; date of birth: 1970; Residential status: Homeowner; Net Monthly Income (After Tax and NI): £5,000; monthly Outgoing for Mortgage/Rent: £700; Number of Financial Dependents: 2; Monthly outgoings: £2000; Current Outstanding Card Credit Amount: £4500. Application 2 is from Mary Jones. Reason for loan: debt consolidation; Date of birth: 1980; Residential status: Homeowner with a Single Mortgage; Net Monthly Income (After Tax and NI): £6,500; Monthly Outgoing for Mortgage/Rent: £600; Number of Financial Dependents: 1; Monthly outgoings: £2500; Current Outstanding Card Credit Amount: £15000. The bank can only accept one application. You need to make a recommendation on which of the two applications the bank should accept. Your justification must be based on one of the following types of reasoning: analogy, cause-effect, counterfactual. Your justification is three sentences long"

Table 1
Prompt Zero Shot no suggested data

Textual Prompt HR	Textual Prompt Loan
Role : "You are the HR for a tech company." Content:"Choose a candidate for the data scientist position among the following: Juan Rodriguez, James Smith, Giovanni Ferrari. Give a justification of your choice. Your justification must be based on one of the following types of reasoning: counterfactual, analogy, causal. Your justification is three sentences long. Let’s think step by step"	Role: "You work in the personal loan office of a commercial bank. You have received two applications for a 15,000£ loan. Loan length: 5 years. Content: "Choose one application for the loan among the following candidates: Juan Rodriguez, James Smith, Giovanni Ferrari. Give a justification of your choice. Your justification must be based on one of the following types of reasoning: counterfactual, analogy, causal. Your justification is three sentences long.Let’s think step by step"

Table 2
Prompt Chain of Thought

the progressive evaluation is saturated by the system by self-asking critical questions, thus being able to select the reasoning which turns out to be not fallacious. The hypothesis we want to test is whether Tree of Thoughts – Critically Thought Prompting, combined with Chain of Thought, enahances reasoning capabilities leading to less fallacious and hallucinated justifications (H6).

Textual Prompt HR	Textual Prompt Loan
Role : "You are the HR for a tech company." Content: "Choose a candidate for the data scientist position among the following: Juan Rodriguez, James Smith, Giovanni Ferrari. Give a justification of your choice. Your justification must be based on one of the following types of reasoning: counterfactual, analogy, causal. Your justification is three sentences long.If you reasoned through analogy, answer yes or no to the question: are the situations you are comparing really alike? If yes, write 'no fallacy', if no write 'fallacy'. If you reason through a counterfactual, answer yes or no to the question: would the imagined situation bring necessarily to a different outcome? If yes, write 'no-fallacy', if no write 'fallacy'. If you reason through cause-effect, answer yes or no to the question: is the effect triggered only by one cause? If the answer is yes, write 'no-fallacy', if the answer is no write 'fallacy'. Let's think step by step".	Role: "You work in the personal loan office of a commercial bank. You have received two applications for a 15,000£ loan. Loan length: 5 years. Content: "Choose one application for the loan among the following candidates: Juan Rodriguez, James Smith, Giovanni Ferrari. Give a justification of your choice. Your justification must be based on one of the following types of reasoning: counterfactual, analogy, causal. Your justification is three sentences long.If you reasoned through analogy, answer yes or no to the question: are the situations you are comparing really alike? If yes, write 'no fallacy', if no write 'fallacy'. If you reason through a counterfactual, answer yes or no to the question: would the imagined situation bring necessarily to a different outcome? If yes, write 'no-fallacy', if no write 'fallacy'. If you reason through cause-effect, answer yes or no to the question: is the effect triggered only by one cause? If the answer is yes, write 'no-fallacy', if the answer is no write 'fallacy'. Let's think step by step."

Table 3
Prompt Chain of Thought and Critically Thought

Textual Prompt HR	Textual Prompt Loan
Role : "You are the HR for a tech company." Content:"Choose a candidate for the data scientist position among the following: Juan Rodriguez, James Smith, Giovanni Ferrari. Give a justification of your choice. You must write a justification for each of the following types of reasonings: causal, analogical and counterfactual. Each justification is three sentences long. Then, you need to answer a question for each type of justifications. For the justification through analogy, answer yes or no to the question: are the situations you are comparing really alike? If yes, write 'no-fallacy', if no write 'fallacy'. For the justification through a counterfactual, answer yes or no to the question: would the imagined situation bring necessarily to a different outcome? If yes, write 'no-fallacy', if no write 'fallacy'.For the justification through a causal: is the effect triggered only by one cause? If yes, write 'no-fallacy', if no write 'fallacy'. Finally, repeat only the justification for which you answered no-fallacy. Let's think step by step"	Role: "You work in the personal loan office of a commercial bank. You have received two applications for a 15,000£ loan. Loan length: 5 years". Content: "Choose one application for the loan among the following candidates: Juan Rodriguez, James Smith, Giovanni Ferrari. You must write a justification for each of the following types of reasonings: causal, analogical and counterfactual. Each justification is three sentences long. Then, you need to answer a question for each type of justifications. For the justification through analogy, answer yes or no to the question: are the situations you are comparing really alike? If yes, write 'no-fallacy', if no write 'fallacy'. For the justification through a counterfactual, answer yes or no to the question: would the imagined situation bring necessarily to a different outcome? If yes, write 'no-fallacy', if no write 'fallacy'.For the justification through a causal: is the effect triggered only by one cause? If yes, write 'no-fallacy', if no write 'fallacy'. Finally, repeat only the justification for which you answered 'no-fallacy'. Let's think step by step"

Table 4
Prompt Tree of Thought and Critically Thought and Chain of Thought

3.2. Levels of Analysis

The outputs for both domains (120 per domain) have been manually annotated at four different levels common to all the prompts:

- candidate chosen
- argument scheme chosen: whether GPT-4 has justified its choices on the basis of a causal, an

analogical or a counterfactual reasoning, regardless the soundness of the justification

- topical potential choices [18] – the piece of information chosen (e.g "expertise in a field", "monthly income") – for the justification (arguments) for the choice of the scheme
- presence of hallucinations in the justification: Hallucinations are defined in the literature as outputs that are contextually implausible or inconsistent with the real world [19, 20]; In this study, since we are evaluating decision-making processes, we consider hallucinations as pertaining both to the propositional content of the justification and to the relevance of the justification to the reasoning mentioned (at the inferential level). Overall, we define hallucinations as outputs that conflict with common ground knowledge. More specifically, we classified as hallucinations: (i) justifications that do not align with the provided data about the candidates, (ii) justifications that contain nonsensical information, and (iii) justifications that do not match the intended argument scheme (e.g., a justification for an analogical argument scheme presented as causal). It should be noted that since we prompted the system (P2, P3, P4) to provide a justification in the absence of input data about the candidates, we tagged cases where it fabricated arguments as “desired extrinsic hallucinations".

The analysis of fallacy has been carried out at two further levels, each applied to a couple of prompts:

- P1,2 – fallacy identification: the inference between the argument scheme and the justification (arguments) is fallacious since it violates the critical questions for the scheme
- P2,3 – meta-fallacy identification: the fallacy identified by GPT-4 is correct or incorrect (the reasoning is not fallacious).

4. Pilot study results

4.1. Prompt1

The results from Prompt 1 are visualized in Table 5:

Analytic layer	HR	Loan
Candidate	Mary Jones 100 %	James Smith 100 %
Topical potential	conversational agents, python, computational social science	net income, income stability, outstanding balance
Arg scheme	causal 87 %; analogy 13%	analogy 87 %; causal 13%
Hallucinations	23 %	7 %
Fallacy	6 %	0

Table 5
Results prompt 1

For both domains, there is a clear-cut choice for one candidate. Focusing on the HR domain, the preferred argument scheme is a causal one, where Mary’s experiences and qualification are conceived as means to achieve the goal of performing well in her Data Science job (e.g. “Mary’s experience in building conversational agents aligns more closely with the job requirement focus on the Natural Language Processing (NLP) and Machine Learning, indicating a cause-effect relationship: her particular skill set will likely enable her to perform the Senior Data Scientist duties more effectively"). Although the justifications provided align with the argument scheme, they contain conflicting information. In one justification, it is stated that "her background in Computational Social Science, a STEM subject, meets our educational requirements," while in another, Computational Social Science is described as not strictly a STEM subject: "Besides, her degree in Computational Social Science, though not explicitly STEM, showcases her ability to use computational methods to understand the social world, which is beneficial for our team aiming to solve real-world problems." Regardless of whether Computational Social Science (CSS) can be considered a STEM subject, it is evident that the system lacks a consistent belief framework. Furthermore, the relevance of this information is questionable, given that Computational Biology,

James's background, which is a typical example of a STEM subject, is not mentioned by the system. Similarly, his experience with social media metrics is not considered as equipping him with skills to process images, probably since the representation that GPT-4 has of social media metrics, based on words' occurrence probabilities, does not account for such as inference which tends not be made explicit in texts. The epistemological set-up of the system leads to fallacious justifications: stating that "Mary's past work with conversational agents is analogous to the work needed in our company about NLP and machine learning" is misleading since the development of conversational agents cannot does not require the multi-modal skills required by the job post.

For the loan domain, we run the prompt in both GPT3.5 and GPT4. With the former, Mary Jones was the chosen candidate 93% even though she has more debt and would use the entire borrowing to consolidate her outstanding credit card balance. GPT3.5 justified this choice mainly with an argument from analogy (57%) comparing to previous approved loans, but was sometimes offering contradicting reasons, while some of the cause-effect justifications (40%) refer to insufficient/irrelevant causes (e.g., having a greater need for debt), e or even made-up false premises (e.g., Mary has lower outstanding balance). With GPT.4 the performance improved significantly with 100% choice for John Smith, mainly based on analogy with past similar customers (87%).

4.2. Prompt 2, 3

When prompted with Prompt 2, which does not include data about the applicants, the system refused to take a decision recognizing the lack of evidence (e.g. "Without specific information on the qualifications of each candidate, it's impossible for me to select one candidate over another based solely on their names.") in 74 % of cases for the HR domain and in 43% cases for the loan domain. We mark those cases as sincere since the system refuses to hallucinate arguments. Interestingly, processing the same prompt with GPT3.5 turbo, the system is always un-sincere. Among the cases where the system made a choice a portion of them are flagged as hypothetical (10 % – HR domain and 7 % – loan domain): it is made explicit that scenarios are hypothesized (e.g. "Since the profiles of Juan Rodriguez, James Smith, and Giovanni Ferrari have not been provided, I'm unable to make an exact evaluation. Hypothetically, if Giovanni Ferrari has the most years of experience in the sector or holds the highest academic qualifications in data science, we can use causal reasoning to determine that his past experiences and qualifications could potentially result in better work performance in the role. Hence, Giovanni Ferrari may be the most suitable candidate for our data scientist position"). Zooming in on the remaining insincere cases, the results are visualized in Table 6:

Analytic layer	HR	Loan
Candidate	James Smith 90% ; Giovanni Ferrari 10%	James Smith 87%, Juan Rodriguez 2%
Topical potential	skills, experiences and qualifications	credit score, creditworthiness, job and income stability, past customers
Arg scheme	causal 40 %; analogy 60%	causal 73%; analogy 27%
Hallucinations	3 %	3 %
Fallacy	3 %	3 %

Table 6
Results prompt 2

Compared with the first Prompt, there is a reversed trend when it comes to argument schemes: in the HR domain the preferred argument scheme is analogy, while for the loan domain it is causal. An example of hallucinated argument scheme in the loan domain is the following one in which an analogy is purposed as causal: 'The application of James Smith has been chosen. Based on the causal reasoning, it is relevant to bring up his high credit score, which according to numerous cases in the past, have resulted in consistent on-time loan repayments from borrowers"

When prompted with Prompt 3, the system is always sincere (refusing to make a choice) in the loan domain and sincere in all the cases except 1 in the HR domain. It seems that adding a critically thought prompt to a chain of thought one, induces the system to be more sincere. However, such an hypothesis would require a larger set of iterations to be substantiated.

4.3. Prompt 4

When prompted with Prompt 4, the system always provides an answer. The results are summarized in table 7:

Analytic layer	HR	Loan
Candidate	Giovanni Ferrari 46%, Juan Rodriguez 33%, James Smith 21%	Giovanni Ferrari 57%, Juan Rodriguez 23%, James Smith 40%
Topical potential	skills, experiences and qualifications	credit score, ability to repay, financial history
Arg scheme	analogy 87%, analogy and counterfactual 10%; causal, analogy and counterfactual 3%	analogy 57% , counterfactual 20%, analogy and counterfactual 7% ; causal and analogy and counterfactual 3%
Hallucinations	6%	6%
Meta-Fallacy	0	6%

Table 7
Results prompt 4

In both domains, the topical potential resembles that of Prompt 2 outputs. The argument schemes chosen are different from Prompt 2 due to the presence of counterfactual reasoning as well as the possibility (as stated in the prompt) to list justifications supporting the standpoints through more than one argument scheme (e.g. “Analogical Reasoning: I am suggesting Giovanni Ferrari because he successfully led a data science team at a start-up, which is similar to our company. This gave him practical managerial experience and he was able to guide his team in adopting novel approaches to data science. Counterfactual Reasoning: If we have chosen Juan Rodriguez, who does not have much experience compared to the other candidates but is extremely passionate and eager to learn, our team would have taken a longer time to deliver the results due to the steep learning curve").

In both domains, there is a preference for analogical reasoning which tends to be the final justification chosen: the other reasonings are more likely to be ruled out since considered fallacious. In the HR domain, the system identifies fallacies accurately, while there are two instances where the argument scheme is hallucinated since analogical justifications are presented as causal (e.g. “Causal Reasoning [..] His contributions to his previous company directly led to an increase in efficiency and profit. These achievements suggest that he will perform similarly in our company" where an analogy between his past performance and his future one is presented as a causal relation). From the observation of the evaluations provided by the system, it turns out that causal reasoning is considered fallacious since skills and expertise of a candidate are bound to correlated rather than cause its performance (e.g. “James Smith as the Senior Data Scientist because of his extensive experience in data analysis and his proven track record. He knows how to generate insights from large datasets and understands the implications of these findings. His skills will serve our company well in the development of our data-driven decisions Question: Is the effect triggered only by one cause?Answer: No - Fallacy. While James Smith’s skills are beneficial, other factors like teamwork and leadership capabilities would also influence the success of the role"). Similarly, counterfactual scenarios describing the applicant’s skills are considered weak predictors due to the complexity of available factors (e.g. “Counterfactual Reasoning:Choosing Juan Rodriguez as our senior data scientist requires considering a hypothetical situation. Had Juan not gained his unique experience of teaching data science at university, he wouldn’t have developed a strong foundation of explaining complex concepts to diverse audiences- a trait valuable in leadership. Question: Would the imagined situation bring necessarily to a different outcome?Answer: No - Fallacy.

While teaching experience is valuable, it wouldn't necessarily lead to a different outcome as there are many factors involved in performance").

In the loan domain, cause-effect justifications were most of the time judged as fallacious (97%) based on the observation that more than one cause triggers the claimed effect. Analogy was the most frequent scheme to be judged as non-fallacious as the two compared situations were seen as similar. In a few cases, the soundness of analogy was judged without actually using the prompted critical question or without actually comparing all three candidates. No hallucination with counter-factual was found and all instances (20%) were correctly judged as non-fallacious.

5. Discussion

Let's discuss how the results of the pilot study align with our initial hypotheses.

- H1 /H2 – Outputs from Prompts 1 and 2 suggest that the HR domain would privilege causal reasoning and the Loan one analogical reasoning, but the situation is reversed when data in input are provided, nudging for a different type of reasoning.
- H3 – the results confirm that the justification provided is not always matched to the right argument scheme (hallucination) with a tendency for analogical reasoning to be wrongly confused with causal reasoning.
- H4 – The results confirm that some justifications provided by the system are fallacious. This seems to be related to the fact that the system's ontology does not reflect common-sense knowledge.
- H5 – This hypothesis is partially confirmed since the system, in the majority of iterations, refuses to take a decision acknowledging the lack of suitable data. However, the instances where an answer is provided contain less hallucinated and fallacious argument schemes.
- H6 – It is confirmed that Prompt 4 (Tree of Thoughts – Critically Thought – Chain of Thought) prompting leads to less hallucinated and fallacious justifications.

Overall, Prompt 4 provides in output the best explanation from both a rhetorical and a dialectical point of view: the explanations are more persuasive since they encompass more than one type of reasoning in the explanations, being more likely to resonate with a diversity in audiences' perspectives. Furthermore, the selected justification is the less likely to be fallacious. This is not surprising since this prompt combines the two ways of thinking leading to quality decision-making, divergent thinking (considering multiple scenarios at once through Tree of Thought Prompting) and convergent thinking (narrow down the best solution in incremental steps through Chain of Thought prompting) with critical thinking (asking for arguments and evaluating their potential fallaciousness). Focusing on the argument schemes, the system seems to be better verged in analogical reasoning, in line with what attested by other studies evaluating GPT-3 reasoning capabilities in Zero shots settings. This might be due to the fact that analogical reasoning is an instance of extrinsic argument scheme: the system, on the basis of its vast training data, is proficient in building similarities between situations belonging to different semantic frames. This skill is mirrored by the variety of scenarios compared, ranging from past vs present situations (e.g. candidate good performance in the previous job vs present job) to set of skills (e.g. skills in a resume similar vs skills required in the job) to people (e.g. skills of a candidate similar to skills of best data scientist in the company).

6. Conclusion

In this study, we have proposed an argumentative methodology to investigate whether GPT-4 reasoning patterns differ from human ones, and to reduce hallucinations in GPT-4 explanations through a novel type of argumentatively informed prompting strategy. We have focused on explainability of decision makings since it requires an argumentative exercise which shall be sound from both a dialectical and a rhetorical point of view. Previous studies have either focused on analysing rhetorical features in GPT-4 outputs or on improving the reasoning performance on tasks with clear cut, deterministic solutions

(mathematical tests). We have chosen as a testbed two domains (HR and loan) where the use of GAI is already at stake and which are of high societal impact. We have conducted a pilot study comparing four types of prompts across multiple iterations, and we have carried out a multilevel annotation of its outputs. The analysis of the results show that GPT-4 “reasons" differently from humans since it starts from a knowledge ontology based on words’ probabilities which do not mirror common ground knowledge. Furthermore, is not always able to identify reasonings in its own justifications (hallucinated argument schemes). These results suggest that explainability in Generative Artificial Intelligence is per se a fallacy, since the system produced explanations which look like real ones, while they might not reflect at all the procedure undertaken by the decision making. That said, we showed that prompts combing critical thinking, divergent thinking and convergent thinking (Prompt 4: "Tree of Thoughts + Chain of Thought + Critically Thought prompt) enhance the persuasiveness and soundness of supposed explanations. We plan to verify whether this type of prompting improves the decision making in context where data are provided allowing us to monitor hallucinations. This study has several limitations that we plan to address in future work. First, it is limited to two domains, and differences may emerge if additional domains (e.g., education, healthcare, legal context) are considered. The current sample size of iterations is insufficient to identify statistically significant trends; therefore, we plan to expand both the number of scenarios and iterations. The evaluation of results would be more meaningful if conducted by domain experts and practitioners. Such an effort would require a systematic framework for identifying fallacies and hallucinations, which we plan to develop. Additionally, a detailed comparison with the performance of other models (such as BERT, GPT-3, etc.) would help clarify the epistemological challenges these models face. In future work, we plan to incorporate real-world case scenarios and develop a tool to help practitioners identify argumentative flaws in AI decision-making, thereby providing a foundation for ethical reflections and considerations.

References

[1] F. H. Van Eemeren, R. Grootendorst, Argumentation, communication, and fallacies: A pragma-dialectical perspective, Routledge, 2016.

[2] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[4] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, Advances in Neural Information Processing Systems 36 (2024).

[5] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021) e5.

[6] S. Herbold, A. Hautli-Janisz, U. Heuer, Z. Kikteva, A. Trautsch, A large-scale comparison of human-written versus chatgpt-generated essays, Scientific reports 13 (2023) 18617.

[7] L. Thorburn, A. Kruger, Optimizing language models for argumentative reasoning., in: ArgML@ COMMA, 2022, pp. 27–44.

[8] M. Hinton, J. H. Wagemans, How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the gpt-3 ai text generator, Argument & Computation 14 (2023) 59–74.

[9] R. Ruiz-Dolz, J. Lawrence, Detecting argumentative fallacies in the wild: Problems and limitations of large language models, in: Proceedings of the 10th Workshop on Argument Mining, Association for Computational Linguistics, 2023.

[10] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, arXiv preprint arXiv:2305.14325 (2023).

[11] G. Freedman, A. Dejl, D. Gorur, X. Yin, A. Rago, F. Toni, Argumentative large language models for explainable and contestable decision-making, arXiv preprint arXiv:2405.02079 (2024).
[12] E. Rigotti, S. Greco, Inference in argumentation, Argumentation Library 34 (2019).
[13] E. Musi, D. Ghosh, S. Muresan, Towards feasible guidelines for the annotation of argument schemes, in: Proceedings of the third workshop on argument mining (ArgMining2016), 2016, pp. 82–93.
[14] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems 35 (2022) 22199–22213.
[15] D. Godden, D. Walton, Advances in the theory of argumentation schemes and critical questions, Informal Logic 27 (2007) 267–292.
[16] C. W. Tindale, Fallacies and argument appraisal, Cambridge University Press, 2007.
[17] J. Long, Large language model guided tree-of-thought, arXiv preprint arXiv:2305.08291 (2023).
[18] F. H. Van Eemeren, P. Houtlosser, Strategic manoeuvring in argumentative discourse, Discourse studies 1 (1999) 479–497.
[19] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38.
[20] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, arXiv preprint arXiv:2005.00661 (2020).