Evaluating Large Language Models on Qualitative Reasoning Tasks: A Case Study using OpenAI's GPT Models

Najwa AlGhamdi*, Kwabena Nuamah and Alan Bundy

Artificial Intelligence and its Applications Institute, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, Scotland

Abstract

This study evaluates the performance of Large Language Models (LLMs) on qualitative reasoning tasks, focusing on identifying inconsistencies in their reasoning processes. Specifically, we examine two versions of OpenAI's General-Purpose Transformer (GPT) models: GPT-3.5 and GPT-4. We hypothesize that LLMs may produce inconsistencies in handling qualitative information, in particular they can generate explanations of their answers that are not faithful to the question. To test this hypothesis, we prompt the models with different scenarios to answer questions from the QuaRTz [1], requiring the models to select the correct answer from options A or B and provide a corresponding explanation for their reasoning process. The evaluation focuses on three key dimensions: the consistency of responses, the faithfulness to the context and the effect of different prompts on their explanation. The results reveal several inconsistencies in the models' explanations of their answers, thereby highlighting challenges in the consistency of their qualitative reasoning.

Keywords

Inconsistency, Faithfulness, Explanation, LLM

1. Introduction

In recent years, the field of natural language understanding (NLU) in reasoning has seen significant advances through the development of large language models (LLMs). LLMs are artificial intelligence models that leverage "large-scale, pre-trained, statistical language models based on neural networks" [2]. Among these models, GPT models have gained substantial attention for their ability to generate human-like text. However, a crucial and challenging aspect of LLMs is explainability, or interpretability, which refers to their ability to understand and describe how a model makes its decisions or arrives at its reasoning [3]. In this study, we use the term *explanation* to refer to the natural text generated by the model to justify its answers.

Generating explanations can sometimes lead to *hallucinations*, where the explanation seems reasonable and logical but is actually meaningless or unfaithful to the context [4, 5]. This issue is related to the faithfulness of the explanation, which concerns how accurately the explanation reflects the model's actual reasoning process [6, 7]. The main challenge is ensuring that the explanations provided by the LLMs are not only plausible but also truly faithful to the model's reasoning. Often, generated explanations do not align with how the model arrived at its conclusions, leading to inconsistencies. For example, consider the following question from the QuaRTz dataset [1]:

> "*Mark studies rocks for a living. He is curious if rocks deeper in the Earth will be warmer or cooler. If Mark were to take the temperature of a rock that is half a mile into the Earth's surface, and then take the temperature of a rock 1 mile into the Earth's surface, he would find that the rock half a mile into the Earth's temperature is... (A) lower (B) higher*"

ABERDEEN'24: Reasoning, Explanations and Applications of Large Language Models, October 17, 2024, ABERDEEN, Scotland
*Corresponding author.
n.alghamdi@sms.ed.ac.uk (N. AlGhamdi); k.nuamah@ed.ac.uk (K. Nuamah); a.bundy@ed.ac.uk (A. Bundy)
0000-0002-9058-4155 (N. AlGhamdi); 0000-0002-6868-9858 (K. Nuamah); 0000-0000-0000-0000 (A. Bundy)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3822/short4.pdf

The model answer was "*(B) higher*" and generated the explanation "*The rock at 1 mile into the Earth's surface will have a higher temperature compared to the rock at half a mile into the Earth as the temperature increases with depth inside the Earth due to geothermal gradient*". In this case, the model correctly explains that temperature increases with depth inside the Earth but then incorrectly answers a question by stating that the deeper rock would be higher (option B), which contradicts its explanation. This example illustrates the importance of clearly understanding the model's reasoning process to identify any inconsistencies in the explanations it generates when justifying its answers. To investigate these issues, we evaluated OpenAI's GPT-3.5 [8] and GPT-4 [9], focusing on their performance in qualitative reasoning tasks using qualitative questions from QuaRTz [1]. In evaluating explanations generated by LLMs, two key dimensions are essential: consistency, and faithfulness.

Consistency refers to the model's ability to provide explanations that do not contradict each other, maintaining consistent answers across different questions [10]. Whereas,

Faithfulness measures the degree to which the explanation accurately represents the model's actual reasoning process. Faithfulness focuses on the transparency and accuracy of the model's reasoning rather than on how convincing the explanation appears to be [7].

These dimensions provide a framework for assessing the quality and reliability of explanations produced by LLMs along with different scenarios of prompting.

The paper is organized as follows. Section 2, gives an overview of the models used in the experiment and the evaluation approach with the dataset used in the evaluation. Section 3, discusses the findings of the experiment. Section 4 briefly discusses some similar work in this area. Finally, in section 5, we conclude the paper with a brief summary and discuss possible future work.

2. Experimental Setup

2.1. Models and Prompting Strategies

This study investigates the performance of OpenAI's GPT models, specifically GPT-3.5-Turbo[8] and GPT-4-Turbo[9], in qualitative reasoning tasks. The primary focus is on identifying inconsistencies between the generated answers and the generated explanations (reasoning processes). To explore this, we employed zero-shot prompting, where the models were only provided with task descriptions without examples and tried different prompts, discussed further in 2.3.

2.2. Dataset

The QuaRTz dataset [1], used in this study, contains open-domain qualitative relationship questions, each presents two possible answers, labelled A and B. In addition to the answer choices, the dataset includes a correct answer key as a ground truth, which is not available to the LLM and is used for assessing the LLM's answers, but does not provide explanations for the correct answers. Each question in the dataset contains contexts or annotations (*para-anno* and *question-anno*) and knowledge statements (*para*), which support the correct answer yet do not leak the correct answer to the LLM explicitly.

This supporting information is part of the evaluating models' understanding of the questions and will be referred to as annotation. Using this dataset will require the model to use multi-hop reasoning and use its background knowledge to handle different types of question. An example of such a question, along with its annotations or context that illustrate how the model constructs its answers, is shown below:

{"answerKey":"A",

"para_id":"QRSent-10162",

"id":"QRQA-10162-3-flip",

[1]The dataset can be found in https://www.kaggle.com/datasets/thedevastator/quartz-a-dataset-of-open-domain-qualitative-rela

”question”:”stem”:”We are designing a submarine to study fish that live far below the surface of the ocean. Before we can send a human researcher down in the submarine, we have to be sure it can tolerate the pressure of the water without cracking. The easier test will be to send our submarine down to”,

”choices”:

[”label”:”A”,”text”:”500 feet or”,

”label”:”B”,”text”:”1500 feet?”],

”para”:”A fluid exerts pressure in all directions, but the pressure is greater at greater depth.”,

”para_anno”:

”effect_dir_sign”:”MORE”,

”cause_dir_sign”:”MORE”,

”effect_prop”:”pressure”,

”cause_prop”:”depth”,

”cause_dir_str”:”greater”,

”effect_dir_str”:”greater”,

”question_anno”:

”less_cause_dir”:”500 feet”,

”more_cause_dir”:”1500 feet”,

”less_effect_prop”:”test”,

”less_effect_dir”:”easier”}

2.3. Evaluation Metrics

The evaluation of the models was centered on three dimensions: consistency, faithfulness of the generated explanations and prompting. First, a manual review was conducted to assess whether the explanations provided by the models were consistent and correctly justified the answers, noting any inconsistencies and instances in which the models failed to generate explanations. Second, the LLM's accuracy was measured by comparing the number of correctly generated answers to the total number of questions, under two instruction sequences in the prompts: one which asked to choose then explain and another asked to explain then choose. This approach allowed us to identify challenges in the consistency of the models' qualitative reasoning and understand how the order of instructions affects the models' responses. Additionally, we used zero-shot prompting which is a technique that instructs LLMs, in this case GPT-3.5 and GPT-4, to use their text generalization capabilities to perform qualitative reasoning tasks without prior training or examples. The models respond to prompts based on their pre-trained knowledge and reasoning skills.

The order of instructions in the prompt and the context, when included in the prompt, can vary and possibly lead to different responses. In this study, we explored two different instruction sequences in our prompts and an annotation. Below is an example of zero-shot prompting with the instructions used and the annotation added:

First, **Choose then Explain:**

> **Prompt:** *“answer the question by choosing A or B then justify the right choice. return A or B in JSON form with fields ‘FinalAnswer’ and a justification field ‘Explanation’ and do not include markdown and write the results in results.jsonl file in JSON form”*

Second, **Explain then Choose (reverse prompt):**

Prompt: *"justify the right choice then answer the question by choosing A or B. return A or B in JSON form with fields 'FinalAnswer' and a justification field 'Explanation' and do not include markdown and write the results in reverseresults.jsonl file in JSON form"*

Third, **Annotation in prompt:**

Prompt: *"the context and question annotations added beside the question and choices"*

3. Findings

3.1. General Observation

The models struggled with maintaining consistency based on random sample of 800 questions from results, especially GPT-3.5, leading to frequent disagreements between the reasoning and the final answer/explanation. Both models could generate plausible explanations, but these did not always imply the correct conclusions due to faulty reasoning or misunderstood concepts. Faithfulness, or the alignment of the explanation with the problem's context, was more reliably maintained by GPT-4. However, the models sometimes generated irrelevant details that mislead the reasoning process[2].

Despite improvements in different scenarios such as adding annotation or changing instruction orders, both models still struggled with handling multi-step reasoning, subtle characteristics, and complex causal relationships. The observed inconsistencies confirm ongoing challenges in achieving human-like reasoning, especially in scenarios requiring detailed contextual and scientific understanding.

3.2. Challenges in Consistency of Explanation in Qualitative Reasoning Based on Manual Evaluation

A key focus of this study was to identify inconsistencies in the models' reasoning processes and it can be found in the generated answers and explanation. table 1

Table 1
The number of answers and explanations that models GPT-3.5 and GPT-4 generated for both scenarios (with and without annotation)

Model (both scenarios)	have explanation	no explanation	have answers	no answers
GPT-3.5	1783	911	2694	0
GPT-4	2694	0	1525	169
Reverse Prompting (GPT-3.5)	2621	73	2694	0

Inaccurate responses were analyzed to determine whether the inconsistencies arose from errors in the reasoning process or misinterpretation of the concepts. Across both models in both scenarios, certain challenges were clear, particularly in handling complex multi-step reasoning or scenarios requiring deep contextual or conceptual understanding. Inconsistent explanations often fall into the following categories with an approximate number of each that appears in both models shown in table 1:

- Misinterpreting causal relationships: the model understood the intent of the question but incorrectly identified/applied causal relationships. For example, in question *"If greenhouse gases in the atmosphere were to be doubled, Earth will become... (A) cooler (B) warmer"*, the model incorrectly provided the explanation: *"Increasing the amount of greenhouse gases should theoretically block more heat from reaching the surface, leading to a cooler planet."*, the response incorrectly applied the causal relationship between greenhouse gases and Earth's temperature.
- Adding irrelevant details: Explanations contains unrelated details to the answer rather than addressing the core question. For instance, the question: *"Causing a series to lose resistors will*

[2]The supplementary materials can be found in https://github.com/NajwaAG/EvaluatingLLM.git

result in a/an .... in the total resistance of the circuit. (A) increase (B) decrease", the explanation given was: *"Adding more resistors in a series circuit increases the total resistance. Conversely, losing resistors, as stated in the question, means reducing the number of resistors, which implies that the total resistance of the circuit will decrease. However, the annotations and the logical relationship between resistors in a series and the total resistance (where more resistors result in an increase in resistance) indicate that the correct context is the effect of having more resistors. Since the question refers to the effect of losing resistors, which is the opposite action, the resulting effect on the total resistance should logically be a decrease. However, the annotations suggest interpreting the increase as the correct outcome possibly due to misunderstanding the question's phrasing or annotations mismatch."*.

- Struggling with precise scientific concepts: Questions involving fine differences in logic or distinctions often lead to inaccurate reasoning. For example, the question: *"When there are fewer carbon atoms in something there are... (A) many options (B) few options"*, led to the explanation: *"When there are fewer carbon atoms in something, there are many options for different compounds and structures to form due to the versatility of carbon bonding"*. The model answered "(A) many options" whereas the correct answer should be "(B) few options," which reflects that fewer carbon atoms would constrain the diversity of compounds rather than expand it.
- Demonstrating the correct concepts, but generating the wrong conclusion: The model understood the underlying principles but misapplied them, leading to an incorrect conclusion. For example, in question: *"We are designing a submarine to study fish that live far below the surface of the ocean. Before we can send a human researcher down in the submarine, we have to be sure it can tolerate the pressure of the water without cracking. The easier test will be to send our submarine down to... (A)500 feet (B)1500 feet"*, the generated explanation was: *"The deeper the submarine goes, the higher the water pressure. Sending the submarine to 1500 feet will provide a more rigorous test of its ability to tolerate pressure compared to sending it to only 500 feet."* the model explained the concept right but failed to deduce the correct conclusion and generate answer "(B) 1500 feet" which is wrong. The model misapplies its correctly understood concept of increasing pressure with depth to conclude that the more depth 1500 feet is the easier test, where the correct answer should be A. 500 feet.

3.3. Accuracy and the Impact of Annotations

When evaluating the models using zero-shot prompting, we found that including annotations had a small impact on accuracy for both GPT-3.5 and GPT-4 as shown in Table 2.

Table 2
Accuracy in percentage and the total number of correct answers out of 2694 questions for both GPT-3.5 and GPT-4 with and without annotation

Model	without annotation		with annotation	
	Accuracy	Correct Answers	Accuracy	Correct Answers
GPT-3.5	73.8%	1993	75.5%	2004
GPT-4	83.3%	2245	83.6%	2252

For GPT-3.5, without annotations, the model correctly answered 1,993 out of 2,694 questions, resulting in an accuracy of 73.8%. With annotations, the accuracy slightly increased to 74.5%, with 2,004 correct answers. This small improvement suggests that while annotations help, they don't make a huge difference in how well the model performs.

GPT-4 performed better overall. Without annotations, it had an accuracy of 83.3%, correctly answering 2,245 out of 2,694 questions. With annotations, the accuracy rose slightly to 83.6%, with 2,252 correct answers. Although GPT-4 was more accurate than GPT-3.5, the benefit of adding annotations was still quite small. These results show that annotations can help improve the performance of large language models on qualitative reasoning tasks, but the improvement is minor. GPT-4 was better at handling reasoning tasks and performed relatively similarly regardless of annotations.

In table 3, a manual evaluation of the results of GPT-3.5 and GPT-4 models, with and without annotations were conducted, which highlights notable differences in handling various types of inconsistencies that are descried in the previous subsection.

Table 3
Number of each type of inconsistencies found in explanation that models GPT-3.5 and GPT-4 generated for both scenarios (with and without annotation)

Model	Irrelevant Details	Scientific Struggles	Contradictions	Misinterpret relationships
GPT-3.5 without anno.	27	15	43	24
GPT-3.5 with anno.	87	34	71	21
GPT-4 without anno.	30	1	26	2
GPT-4 with anno.	95	79	65	5

The table shows that GPT-4 with annotations tends to include too many irrelevant details to its explanations with 95 instances and often misunderstood scientific concepts in 79 questions. In contrast, GPT-4 without annotations has fewer issues with scientific concepts, which suggests that although annotations are intended to enhance the model's understanding and responses, it can make them worse. Based on the results, it probably contributes to complicate the reasoning process and to being unable to answer all questions, even if it generate an explanation. Meanwhile, GPT-3.5 when annotated, shows a significant number of instances where the model correctly understands the concepts of 71 questions but leads it to wrong conclusions and generates incorrect answers, 24 of explanations misinterpret the relationships and more irrelevant details appears in 87 explanation which is more comparing to same model without annotation, which was 27 instances. Additionally, GPT-3.5 without annotation shows 15 instances where it struggles with precise scientific concepts which is less than when the same model is: 43 occurrences where it understands the correct concept but concludes incorrectly, and 24 cases of misinterpreting relationships. This experiment is limited by the use of a closed model, and the manual evaluation may be subject to subjective biases. These limitations could be overcome by using other open LLMs such as LLaMA [11] and by involving a larger number of participants in the manual assessment.

4. Related work

Large Language Models (LLMs), like GPT-3.5 and GPT-4, have significantly impacted the field of natural language understanding (NLU) in various reasoning tasks, including translation and question-answering[12, 13]. These models have shown noticeable ability in generating human-like text and dealing with reasoning tasks[12]. However, one of the ongoing challenges is the explainability of these models, i.e., how they generate responses and whether these responses are based on logical reasoning. Inconsistencies and hallucinations found in explanations provided by models can sometimes seems reasonable and logical but are meaningless or unfaithful to the context [4, 5, 14]

As pointed by [15] and [16] LLMs often raise challenges regarding the transparency of their decision-making/reasoning processes and it can be clearly seen in the explanation of their answers. Many studies evaluate the models by focusing on only the annotations/context added in the prompt and how they reflect the generate related explanation to the question [17, 18, 19].Additionally, [20] identifies two primary sources of hallucination in three language models (LLaMA, GPT-3.5, and PaLM): first, the models may assert a logical connection based on the presence of similar statements in their training data, even if these are irrelevant to the current context; second, the models may make incorrect assumptions based on similar frequency patterns rather than logical reasoning. The study shows that while LLMs can generate plausible responses, these are often not grounded in a true the context understanding but rather rely on memorized data. The study addresses these issues by isolating the effects of these hallucination on model performance. The paper [21] evaluates how LLMs like Flan-T5, Alpaca, GPT-3.5, and Llama-2 handle prompting, focusing on identifying two types of inconsistencies: the irrelevant

details and the generation of contradictory or factually incorrect responses. The study utilizes human assessments to evaluate the models' correctness and faithfulness to the provided information. This approach aims to enhance the models' capabilities in producing accurate and contextually appropriate responses.

Compared to prior work, our study evaluated explanations provided along with different prompting scenarios and how they affect each other, then identify any inconsistencies from qualitative questions and how many times it appears, which is not included in previous works.

5. Conclusion and Future work

In this study, we evaluated the performance of OpenAI's GPT-3.5 and GPT-4 models on qualitative reasoning tasks, with a focus on understanding how well these models generate explanations that are consistent and faithful to their reasoning processes. Our findings indicate that while GPT-4 performs better in the view of overall accuracy and consistency compared to GPT-3.5, both models still face challenges, particularly in handling complex reasoning tasks that require a deep understanding of context and scientific concepts. Issues such as misinterpretation of causal relationships and the overemphasis on irrelevant details highlight the limitations of these models in producing reliable, human-like reasoning.

Future work may focus on improving how models deal with information overload, which seems to hurt their reasoning abilities. We should explore ways to help models handle multi-step reasoning better. Additionally, we will look at using few-shot prompting—where we give an example in the prompt—to see how it affects model performance. It's also crucial to study how information overload relates to reasoning accuracy. Further research is needed to make sure the explanations from large language models are consistent and trustworthy. This will help ensure that their answers are not just convincing but also truly reflect how they arrive at their conclusions.

References

[1] O. Tafjord, M. Gardner, K. Lin, P. Clark, QuaRTz: An open-domain dataset of qualitative relationship questions, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5941–5946. URL: https://aclanthology.org/D19-1608. doi:10.18653/v1/D19-1608.

[2] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, arXiv preprint arXiv:2402.06196 (2024).

[3] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, M. Du, Explainability for large language models: A survey, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–38.

[4] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al., Siren's song in the ai ocean: a survey on hallucination in large language models, arXiv preprint arXiv:2309.01219 (2023).

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

[6] A. Jacovi, Y. Goldberg, Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?, arXiv preprint arXiv:2004.03685 (2020).

[7] S. Wiegreffe, A. Marasović, Teach me to explain: A review of datasets for explainable natural language processing, arXiv preprint arXiv:2102.12060 (2021).

[8] OpenAI, gpt-3-5-turbo, 2023. URL: https://platform.openai.com/docs/models/gpt-3-5-turbo.

[9] OpenAI, gpt-4-turbo-and-gpt-4, 2023. URL: https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4, accessed on: September 15, 2023.

[10] P. Hase, M. Bansal, Evaluating explainable ai: Which algorithmic explanations help users predict model behavior?, arXiv preprint arXiv:2005.01831 (2020).

[11] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[12] T. B. Brown, Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).
[13] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[14] T. Xue, Z. Wang, Z. Wang, C. Han, P. Yu, H. Ji, Rcot: Detecting and rectifying factual inconsistency in reasoning by reversing chain-of-thought, arXiv preprint arXiv:2305.11499 (2023).
[15] 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, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.
[16] K. McGrath, Unveiling the power and limitations of large language models, 2024. URL: https://www.6clicks.com/resources/blog/unveiling-the-power-of-large-language-models.
[17] A. K. Lampinen, I. Dasgupta, S. C. Chan, K. Matthewson, M. H. Tessler, A. Creswell, J. L. McClelland, J. X. Wang, F. Hill, Can language models learn from explanations in context?, arXiv preprint arXiv:2204.02329 (2022).
[18] S. Teso, Ö. Alkan, W. Stammer, E. Daly, Leveraging explanations in interactive machine learning: An overview, Frontiers in Artificial Intelligence 6 (2023) 1066049.
[19] J. Kunz, M. Kuhlmann, Properties and challenges of llm-generated explanations, arXiv preprint arXiv:2402.10532 (2024).
[20] N. McKenna, T. Li, L. Cheng, M. J. Hosseini, M. Johnson, M. Steedman, Sources of hallucination by large language models on inference tasks, arXiv preprint arXiv:2305.14552 (2023).
[21] V. Adlakha, P. BehnamGhader, X. H. Lu, N. Meade, S. Reddy, Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering, Transactions of the Association for Computational Linguistics 12 (2024) 681–699. URL: https://doi.org/10.1162/tacl_a_00667. doi:10.1162/tacl_a_00667. arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00667/2374800/tacl_a_00667.pdf.