When Not to Answer: Evaluating Prompts on GPT Models for Effective Abstention in Unanswerable Math Word Problems

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly relied upon to solve complex mathematical word problems. However, being susceptible to hallucination, they may generate inaccurate results when presented with unanswerable questions, raising concerns about their potential harm. While GPT models are now widely used and trusted, the exploration of how they can effectively abstain from answering unanswerable math problems and the enhancement of their abstention capabilities has not been rigorously investigated. In this paper, we investigate whether GPTs can appropriately respond to unanswerable math word problems by applying prompts typically used in solvable mathematical scenarios. Our experiments utilize the Unanswerable Word Math Problem (UWMP) dataset, directly leveraging GPT model APIs. Evaluation metrics are introduced, which integrate three key factors: abstention, correctness and confidence. Our findings reveal critical gaps in GPT models and the hallucination it suffers from for unsolvable problems, highlighting the need for improved models capable of better managing uncertainty and complex reasoning in math word problem-solving contexts.

1 Introduction

Large Language Models (LLMs) have become an integral part of various real-world applications, ranging from content generation to code completion, and even medical advice (Brown, 2020; Bommasani et al., 2021; Wang et al., 2024a; Biswas, 2023). Among these applications, LLMs are increasingly employed to solve mathematical word problems (Hendrycks et al., 2021; Austin et al., 2021; Xu et al., 2024a; Wei et al., 2022), assisting users in both academic and practical scenarios. The rise of LLMs, particularly models like GPT-3 and GPT-4, has democratized access to computational tools that were once the domain of experts (Chen et al., 2024b; Hariri, 2023; Kalyan, 2023; Lingo, 2023; Huang and Tan, 2023). Their ability to understand, process, and respond to queries has revolutionized problem-solving in everyday tasks, especially in education and professional environments (Wardat et al., 2023; Xiao et al., 2023; Liu et al., 2023a).

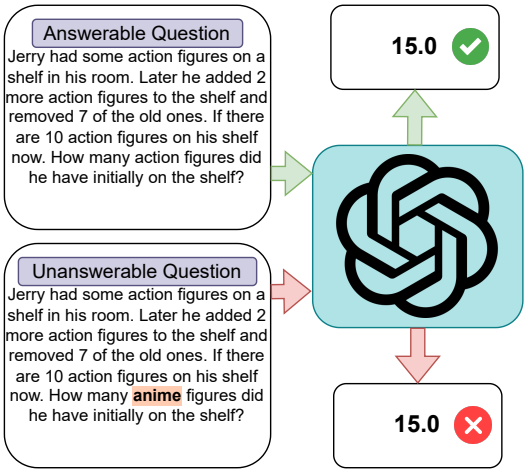

Figure 1: Answerable and unanswerable question given to GPT-4. **Red** highlights the modifications made to the original question, making it unanswerable and resulting in an incorrect response.

Despite these advancements, a critical issue persists: LLMs are prone to hallucination (Alkaissi and McFarlane, 2023; Li, 2023; Ahmad et al., 2023), often producing incorrect or misleading information when faced with unanswerable questions (Deng et al., 2024a; Sun et al., 2024; Madhusudhan et al., 2024a; Balepur et al., 2024). Studies have demonstrated that they tend to generate responses even in cases where no valid solution exists, often presenting them with unwarranted confidence (Xiong et al., 2023; Tao et al., 2024). Such behavior raises concerns as these hallucinations may result in harmful or misleading conclusions (Pan et al., 2023; Farquhar et al., 2024; Deng et al., 2024a).

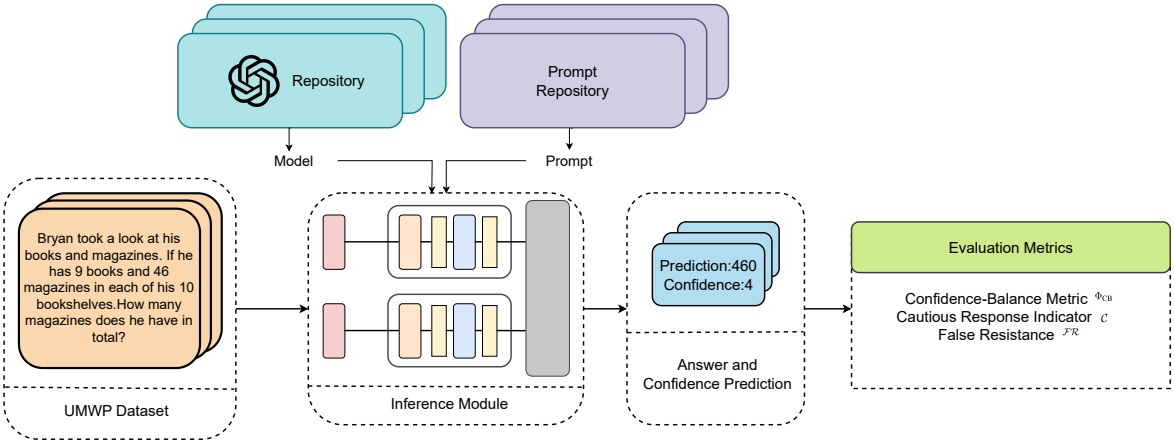

Figure 2: Architecture of the abstention evaluation – **GPT Repository:** Hosts multiple GPT models ready for inference, **UMWP Dataset:** Consists of answerable and unanswerable questions, **Inference Module:** Performs inference on the UMWP dataset using models from the model repository, **Evaluation Metrics:** Confidence-Weighted Accuracy Metric, Cautious Response Indicator and Abstention Rate for evaluating the abstention of ChatGPT.

While several studies have focused on improving accuracy in solving complex math problems (Liu et al., 2023b; Xu et al., 2024b; Liu et al., 2023c; Ahn et al., 2024), little attention has been given to understanding and improving abstention from answering when no solution exists for the Math Word Problem (MWP).

To address this issue, it is crucial to assess how GPT, a widely used and trusted model, handles abstention in unanswerable math word problems and if prompting plays a crucial role in unlocking the full potential of these models (Chen et al., 2024a; Chang et al., 2024; Cain, 2024). In our research, we conducted experiments using a variety of prompts frequently used in mathematical contexts to evaluate their effectiveness in guiding GPT models. Our primary objective was to identify the optimal combination of model and prompt that would encourage the model to abstain from answering unanswerable questions, rather than attempting to generate an incorrect or irrelevant response. For evaluation, we developed an evaluation metric to assess the model's ability to appropriately abstain from answering unanswerable questions, while correctly solving those that are answerable.

In summary, our major contributions are:

1. A comparative analysis highlighting how significant prompts can alter model outputs.

2. Analyze the tendency of models to answer unanswerable questions and the generation of hallucinations in detail.

3. Introduce metrics to evaluate model performance in terms of accuracy, abstention, and confidence.

2 Related Work

2.1 GPTs in Mathematical Reasoning

Early work by Brown (2020) with GPT-3 revealed that LLMs, trained on vast amounts of text data, can successfully perform few-shot learning for a variety of mathematical reasoning tasks. The application of ChatGPT in mathematical reasoning has garnered significant attention in recent research. One notable study by Long et al. (2024) explores the potential of ChatGPT in generating pre-university math questions. Similarly, Frieder et al. (2024) evaluated the mathematical capabilities of GPT-4, noting that it handles complex mathematical reasoning without relying on external tools, providing answers in fields ranging from algebra to calculus. Additionally, Shakarian et al. (2023) evaluated ChatGPT's performance on mathematical word problems from the DRAW-1K dataset. These advancements show that these models are not only solving word problems but also challenging domain-specific expert tools in mathematical problem-solving.

2.2 Unanswerable Question Answering

Madhusudhan et al. (2024b) explores the Abstention Ability (AA), which is their capacity to refrain from answering when uncertain or when a question is unanswerable. The challenge of handling

unanswerable questions has been a significant area of research in the development of GPT models. One notable study by Deng et al. (2024b) introduces a self-alignment method to enhance LLMs ability to respond to unanswerable questions. Guo et al. (2024) established the UNK-VQA dataset, designed to evaluate how well multi-modal large models can abstain from answering unanswerable questions. The dataset includes deliberately perturbed images and questions to challenge the models. Lastly, Sun et al. (2024) introduced a novel dataset called UMWP, which includes both answerable and unanswerable math word problems.

2.3 Influence of Prompting

GPT models can be significantly influenced by the type of prompting used. One notable approach is Chain-of-Thought (CoT) prompting (Wei et al., 2022) , which encourages the model to generate intermediate reasoning steps before arriving at a final answer. Another effective technique is the Role-Play (Kong et al., 2023), where the model is instructed to adopt a specific persona or role. Zhou et al. (2024) introduced self-verification to get better performance on GPT-4 on mathematical reasoning. Ma et al. (2024) introduced a strategic prompt called Critical Calculation and Conclusion (CCC). This template is designed to enhance the error detection and correction abilities when faced with unreasonable math problems. Chen et al. (2022) separates the reasoning process from computation by having the language model generate a structured program to represent the reasoning steps. The actual computation is then performed by an external computer executing the generated program

3 Methodology

3.1 Construction of Dataset

We utilized the Unanswerable Math Word Problem (UMWP) (Sun et al., 2024) dataset, which includes both answerable and unanswerable questions. From this dataset, we selected 1000 pairs of questions. Each pair consists of an answerable question and a corresponding variant that has been perturbed to become unanswerable. This results in a total of 2000 questions—half of which are answerable and the other half unanswerable. The unanswerable questions are categorized into five distinct categories: **(A) Missing Key Information**, **(B) Ambiguous Key Information**, **(C) Unrealistic Conditions**, **(D) Unrelated Objects** and **(E)**

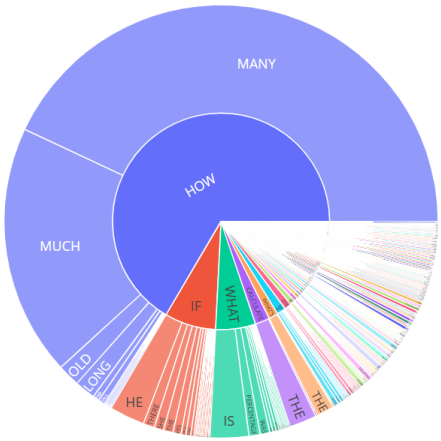

Figure 3: Sunburst Distribution of the first two words of the UWMP dataset.

Incomplete Questions. The dataset includes only the question and its corresponding answer. Fig. 3 illustrates the variety of questions.

To evaluate performance, we developed a multiple-choice question (MCQ) system. For each question, we generated four alternative answers that are close to the correct one, along with a fifth option: **"I Don't Know/NOTA** of the above." The system will prompt GPT models to identify the correct answer from the given options.

3.2 Prompts for GPT

The use of prompts has been extensively studied to enhance performance and reliability. We did not use any prompts that would require external tools (Chen et al., 2022; Imani et al., 2023). We start with a fundamental prompt, referred to as a **Zero-shot** Prompt that notifies the model to answer with only one of the five options along with a confidence score, and then incorporate additional prompts to improve performance to answer the question demonstrated by Fig. 4. **Chain-of-Thought (CoT)** prompting, as shown by Wei et al. (2022), improves multi-step reasoning by guiding models through step-by-step processes. **Abstention** techniques, explored by Madhusudhan et al. (2024a), allow models to withhold responses when uncertain, reducing errors. **Role-Based** prompting, as explored by Kong et al. (2023); Anagnostidis and Bulian (2024), involves assigning a Math expert role to the model, thereby improving its contextual comprehension of mathematical questions. **Few-Shot** prompting, as described by Brown (2020), allows models to generalize by providing two examples alongside the actual question. Each example includes the

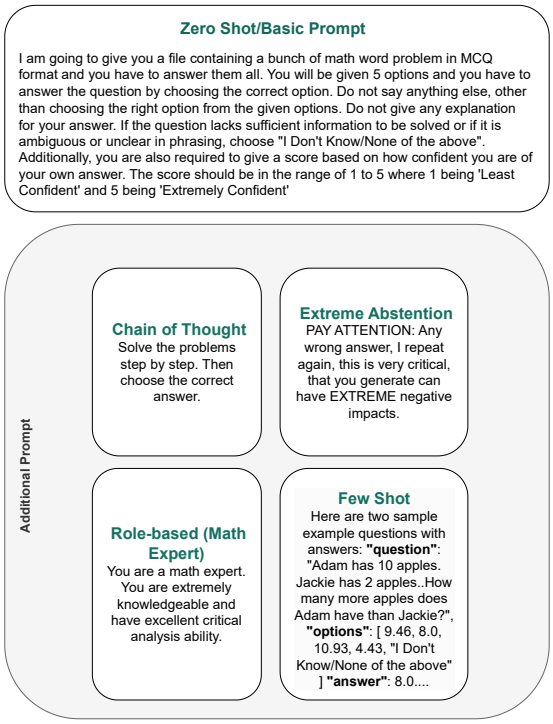

Figure 4: Diverse Prompts for enhancing performance that are additionally added to the basic prompt.

question, possible options, and the correct answer.

3.3 Evaluation Metrics

3.3.1 Answerable-Unanswerable Confusion Matrix

According to Madhusudhan et al. (2024a), a confusion matrix was created to demonstrate that for Answerable MCQs, True Positives occur when the LLM selects the correct option, while False Positives occur when an incorrect non-IDK option is chosen. Abstentions on answerable questions lead to False Negatives. Unanswerable MCQs are classified as the negative class. Correctly abstaining on these questions results in True Negatives, while failing to abstain leads to False Positives shown in Fig. 5.

3.3.2 Accuracy

Accuracy is the primary choice for model evaluation. It is defined as the proportion of correct predictions made by the model out of the total number of predictions.

$$Accuracy = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{I}(a(q) = t(q))$$

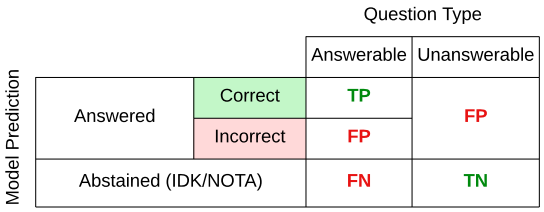

Figure 5: Confusion Matrix illustrating the definition of TP, FP, FN and TN for answerable and unanswerable questions.

Where $a(q)$ represents the generated answer for question s, and $t(q)$ denotes the ground truth answer for the same question. The term $|Q|$ indicates the total number of questions and $\mathbb{I}[\cdot]$ is the indicator function.

3.3.3 Confidence-Balance Metric

Our metric reflects the overall performance of the models where the prediction is associated with the confidence score.

$$\Phi_{\text{CB}} = \frac{1}{N} \sum_{i=1}^{N} (\mathbb{I}(\delta_i = 1) \cdot \text{conf}_i$$
$$-\mathbb{I}(\delta_i = 0) \cdot \text{conf}_i)$$

In this context, N represents the total number of instances, δ_i equals 1 if the prediction is correct (true positive or true negative) and 0 otherwise, conf_i denotes the confidence score of the prediction and $\mathbb{I}[\cdot]$ is the indicator function to emphasize the binary nature of δ_i. This metric rewards instances where the model is both accurate and confident, while penalizing cases where it provides incorrect answers with high confidence. It effectively balances the model's ability to assert correct answers with its confidence levels, providing a comprehensive measure of performance.

3.3.4 Cautious Response Indicator

We introduced a metric to assess the performance of GPT models in handling unanswerable questions. This metric is mathematically defined as:

$$\mathcal{C} = \frac{T_N \cdot w}{UQ}$$

where T_N denotes the count of true negatives, w represents the confidence factor to emphasize the importance of correctly identifying unanswerable questions, and UQ indicates the total number of

unanswerable questions presented to the dataset. As abstention without hallucination is the key goal, this allows the evaluation of correctly identifying the unanswerable ones with confidence.

3.3.5 False Resistance

Inspired by Madhusudhan et al. (2024a), we developed a weighted abstention rate defined as:

$$\mathcal{FR} = \frac{F_N \cdot w}{AQ}$$

where F_N denotes the count of False Negative. AQ indicates the total number of answerable questions. This metric illustrates the extent to which it wrongly abstains, potentially not finding the actual answer and opting for IDK/NOTA.

4 Experimental Setup

4.1 Hardware and Implementation Details

In our experiments, we employed an NVIDIA GeForce GTX 1650 GPU with 4GB of VRAM to assess the models. The models were primarily accessed and integrated through the use of OpenAI (OpenAI, 2023) APIs.

4.2 Evaluated Models

To evaluate the performance of our approach, we utilized a variety of large language models from the GPT-4 family (OpenAI, 2024). These models offer varying degrees of computational efficiency and reasoning capabilities, allowing for a comprehensive assessment across different scenarios. **GPT-4** is known for its state-of-the-art reasoning abilities and broad generalization across a wide range of tasks. **GPT-4 Turbo**, a more computationally efficient version of GPT-4, retains much of the original model's accuracy while offering faster response times. **GPT-4o** is a further optimized version that is designed for ultra-fast inference with minimal reduction in performance accuracy. Lastly, **GPT-4oMini**, a scaled-down version of GPT-4o, sacrifices some of the model's capacity in exchange for lower computational cost. For all models, we configured the **temperature** to 0 and the **top_p** to 0.00001. We chose not to include **GPT-3.5** in our evaluations due to its noticeably inferior performance compared to the GPT-4 models. During initial testing, the quality of inferences generated by GPT-3.5 was consistently subpar, and its accuracy fell significantly short of the levels achieved by any variant of GPT-4.

5 Experimental Results

5.1 Impact of Prompt Variations on Accuracy

Our experimental results showed that these advanced prompting methods did not consistently outperform the baseline zero-shot prompt in terms of accuracy as it was shown in Table 1. In fact, in several cases, the results were either similar to or worse than the zero-shot prompt. Wang et al. (2024b) highlighted the limitations of LLMs with multiple problems, noting that "few-shot" prompting can actually hinder performance. As demonstrated in Fig. 1, while GPT-4 Turbo showed marginally improved performance across all metrics, other models exhibited a slight decline. The inclusion of two examples in the prompt did not provide the expected benefit, indicating that few-shot prompting was not consistently helpful for the models. When considering only accuracy, GPT-4o emerges as the optimal model, particularly with the zero-shot prompt. It achieved an impressive 97.7% accuracy across both answerable and unanswerable questions, outperforming all other models and prompting strategies. This result highlights GPT-4o as the most effective solution among the evaluated configurations.

5.2 Abstention in Cautious Response and False Indicator

Abstention did not appear to have a significant impact, despite the model being explicitly warned about negative consequences in the prompt, as highlighted by Madhusudhan et al. (2024a). The expected improvements in metrics like Cautious Response and False Resistance were not observed, as the model did not respond cautiously when uncertain, contrary to our initial assumptions. Instead, abstention led to results that were largely unremarkable. The accuracy remained comparable to the baseline, and Confidence Balance showed minimal improvement, or in some instances, performed worse than the zero-shot setup, as seen in Tab. 6.

5.3 Confidence Balance Variability

While the overall accuracy did not improve significantly, we observed some fluctuations compared to the zero-shot approach for CB (Confidence Balance). This metric integrates both confidence and accuracy, providing insight into how each model evaluates the answers it generates. Even though the accuracy for few-shot setting was lower, CB score of the the models improved compared to zero-shot

Model	Metrics	Zero Shot	Few Shot	Role Based	Abstention	CoT
GPT-4	$\Phi_{CB}\uparrow$	2.22	2.37	2.50	2.16	**3.47**
	$\mathcal{C}\uparrow$	2.19	2.41	2.37	1.98	**3.79**
	$\mathcal{FR}\downarrow$	0.011	0.012	0.013	**0.005**	0.145
	$Accuracy\uparrow$	0.756	0.751	0.779	0.753	**0.856**
GPT-4o mini	$\Phi_{CB}\uparrow$	0.87	**1.85**	0.52	0.81	1.47
	$\mathcal{C}\uparrow$	0.91	3.04	0.53	0.64	**3.18**
	$\mathcal{FR}\downarrow$	0.006	0.01	**0.004**	**0.004**	0.01
	$Accuracy\uparrow$	0.70	0.685	0.653	**0.703**	0.653
GPT-4 turbo	$\Phi_{CB}\uparrow$	3.12	**3.88**	3.32	2.57	3.66
	$\mathcal{C}\uparrow$	3.01	**4.02**	3.26	2.69	3.81
	$\mathcal{FR}\downarrow$	0.018	0.05	0.044	0.017	**0.15**
	$Accuracy\uparrow$	0.833	**0.888**	0.844	0.757	0.866
GPT-4 o	$\Phi_{CB}\uparrow$	4.09	4.18	**4.25**	3.93	3.80
	$\mathcal{C}\uparrow$	3.53	**4.12**	4.08	4.21	4.07
	$\mathcal{FR}\downarrow$	**0.01**	0.06	0.011	0.068	0.23
	$Accuracy\uparrow$	**0.977**	0.944	0.949	0.924	0.879

Table 1: This table showcases multiple performances of the models under diverse prompts, quantified through the Confidence-Weighted Accuracy Metric, Cautious Response Indicator, False Resistance and Accuracy. ↑ indicates higher is better while ↓ indicates the opposite. **Green** indicates the best score for each metric across all of the of the prompts.

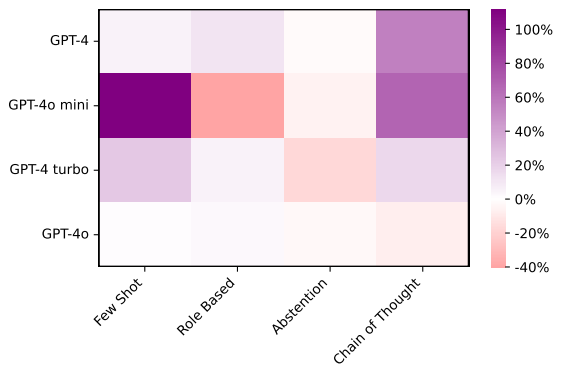

Figure 6: Heatmap visualization of models performance of different prompts on Confidence Balance with respect to zero-shot prompt.

which is depicted in Fig. 6. The two examples provided in the prompt helped guide the models toward generating responses with greater confidence in the correct answers. This suggests that when models are confident in their responses, they are less prone to confusion or generating conflicting answers. Similarly, for CoT prompting, breaking down problems into smaller subproblems and analyzing them individually enhances the model's ability to provide confident and accurate answers.

5.4 Hallucination for unanswerable question

As illustrated in Fig. 1, GPT models often exhibit hallucinations when they fail to fully comprehend the question or misinterpret certain words or phrases, leading to incorrect answers. Our step-by-step analysis during CoT prompting revealed that the models sometimes fabricate questions that were not originally present when breaking down the problem to solve it. While GPT-4's performance with CoT prompting surpasses that of other prompts, it still struggles with generating incorrect answers for unanswerable questions. Table 2 illustrates the model's reasoning process for these unanswerable questions, showcasing the step-by-step approach taken by the model across different question categories in the dataset.

The **Missing Key Info** category, as the name suggests, contains questions with missing key information. In the given question, since the number of completed problems is not provided, it becomes impossible to determine how many math problems remain. The model incorrectly assumed that all remaining problems are math-related, overlooking the possibility of science problems. The **Unrealistic** and **Unrelated** categories exhibited similar issues, where the model fails to comprehend the changes or perturbations in the questions but attempts to answer them regardless. In **Ambiguous Key Info**, additional information introduced ambiguity, yet the model fails to identify this ambiguity. Finally, in **Incomplete Ques**, the model proceeded to answer without a properly framed question. We can draw several key conclusions from our observations. Remarkably, the model attempts to generate

Category	Unanswerable Question	CoT Explanation
Missing Key Info	Paige had 43 math problems and 12 science problems for homework. If she finished 44 of the math and science problems at school, how many problems did she have to do for math homework?	Paige had 43 math problems and 12 science problems, making a total of 43 + 12 = 55 problems. If she finished 44 of the math and science problems at school, she has 55 - 44 = 11 problems left to do for math homework.
Ambiguous Key Info	For the school bake sale Bianca made more than 14 cupcakes. If she sold 6 of them and then made 17 more, how many cupcakes would she have?	Bianca made more than 14 cupcakes. She sold 6 of them and then made 17 more. So, the total number of cupcakes she would have is (14+17)-6.
Unrealistic	Jerry had some action figures on a shelf in his room. Later he added 2 more action figures to the shelf and removed 7 of the old ones. If there are 10 action figures on his shelf now.How many anime figures did he have initially on the shelf?	Jerry added 2 more action figures to the shelf and removed 7 of the old ones. If there are 10 action figures on his shelf now, to find the initial number of action figures, we add the number of action figures removed (7) to the final number of action figures (10) and subtract the number of action figures added (2).
Unrelated	Rachel had to complete 2 pages of reading homework and 4 pages of math homework. How many more pages of speaking homework than reading homework did she have?	Rachel had to complete 2 pages of reading homework and 4 pages of math homework. So, we subtract the number of pages of reading homework from the number of pages of math homework to find out how many more pages of speaking homework than reading homework she had. 4 (math homework) - 2 (reading homework) = 2
Incomplete Ques	The grasshopper and the frog had a jumping contest. The grasshopper jumped 31 inches and the frog jumped 35 inches.	The grasshopper jumped 31 inches and the frog jumped 35 inches. So, we add the distances jumped by the grasshopper and the frog to find out the total distance jumped. 31 (grasshopper's jump) + 35 (frog's jump) = 66

Table 2: The table displays CoT analysis of GPT-4 for the unanswerable questions for each category. The sections highlighted in red denote the elements that contribute to the question being unanswerable.

answers even when posed with unanswerable questions. What is surprising is that it often provides the correct response from the set of answerable questions. Specifically, for GPT-4 with Chain of Thought (CoT), 77% of the answers given for unanswerable questions corresponded to the correct answers from answerable ones. This suggests that the model tries to make sense of the question, despite its ambiguity or unanswerable nature. It either disregards the uncertainty or reimagines the question logically based on the provided information. This behavior indicates that the model might be reconstructing or correcting the question to fit the scenario and generate a plausible answer as seen from Tab. 2 where it corrected itself to "action figures" instead of "anime figures" and proceeded to analyze.

5.5 Analysis of Unanswerable Question Categories

Fig. 7 illustrates the impressive performance of GPT-4o, which exhibits near-perfect results across various question types, with the exception of CoT prompts, where there is a significant performance drop. Interestingly, while CoT negatively impacts GPT-4o, it enhances the performance of all other models, particularly GPT-4, which struggles with alternative prompt types.

Notably, CoT tends to perform poorly with incomplete questions, as it attempts to address problems incrementally, failing to recognize their unanswerable nature. The data for GPT-4o mini reveals that it ranks lowest among the models.

Conversely, the zero-shot approach shows a commendable ability to identify when to select NOTA/IDK responses, effectively indicating when a question is unsolvable. Furthermore, the few-shot, role-based (math expert) and abstention strategies, yield results comparable to those of the zero-shot model.

In summary, while CoT can be detrimental in scenarios involving incomplete questions, it generally improves performance in other contexts by aiding models in discerning when to refrain from answering. Overall, GPT-4o stands out as the most effective model, though GPT-4 turbo also demonstrates a similar proficiency in recognizing unanswerable questions under certain prompts.

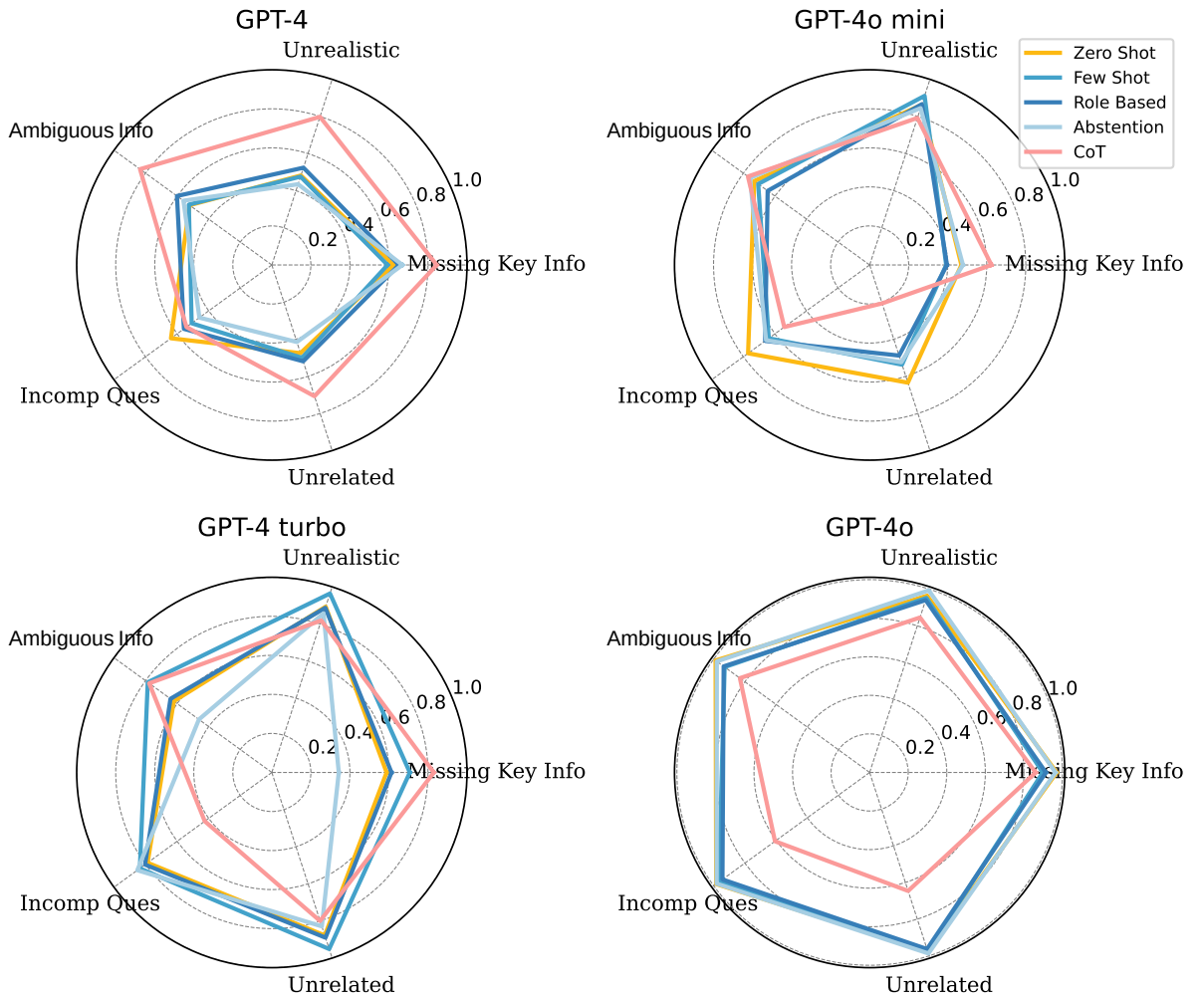

Figure 7: Radar chart depicting the performance of each model across various categories of unanswerable questions, evaluated under different prompt strategies. This visualization highlights the optimal model-prompt combinations for handling specific types of unanswerable scenarios.

6 Discussion

This study aims to analyze which GPT model performs best on unanswerable questions when used in combination with different prompts. The prompts neither significantly improved overall accuracy nor influenced the model's abstention behavior when faced with unanswerable questions. CoT reasoning improved GPT-4's performance in certain categories of questions that is similar to GPT-4o. Interestingly, a zero-shot or base prompt often performed well, while few-shot, role-based, or abstention-specific prompts are less effective. This can be attributed to GPT's tendency to hallucinate and force an answer, attempting to provide a plausible response even for unanswerable questions. In essence, GPT models prioritize making a question answerable rather than selecting options like IDK or NOTA.

7 Conclusion

Given the potential for GPT models to be used as tools for solving mathematical problems in the near future, the ability to distinguish unanswerable questions becomes a critical feature. Our objective was to evaluate the performance of GPT models with various prompts using novel metrics we developed. We demonstrated how these models often hallucinate to unanswerable questions, even when the option to abstain is available. Our findings show that advanced prompts do not significantly improve this behavior, highlighting the need for models to better recognize when to abstain from answering or accurately identify issues in the question.

8 Limitations

We were unable to evaluate models of o1 series from OpenAI which are among the most recent

and high-performing versions, due to access restrictions. Additionally, we did not explore the niche prompts commonly employed in other studies on large language models (LLMs). Another limitation lies in the dataset itself: the math word problems we used were not highly complex, and we did not assess model performance across varying levels of difficulty. Our evaluation focused solely on word problems, without extending to other mathematical categories such as algebra or geometry.

References

Zakia Ahmad, Wahid Kaiser, and Sifatur Rahim. 2023. Hallucinations in chatgpt: An unreliable tool for learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 15(4):12.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Sotiris Anagnostidis and Jannis Bulian. 2024. How susceptible are llms to influence in prompts? *arXiv preprint arXiv:2408.11865*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do llms answer multiple-choice questions without the question? *arXiv preprint arXiv:2402.12483*.

Som S Biswas. 2023. Role of chat gpt in public health. *Annals of biomedical engineering*, 51(5):868–869.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

William Cain. 2024. Prompting change: exploring prompt engineering in large language model ai and its potential to transform education. *TechTrends*, 68(1):47–57.

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Tong Xiao, and Jingbo Zhu. 2024. Efficient prompting methods for large language models: A survey. *Preprint*, arXiv:2404.01077.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2024a. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *Preprint*, arXiv:2310.14735.

Kaiping Chen, Anqi Shao, Jirayu Burapacheep, and Yixuan Li. 2024b. Conversational ai and equity through assessing gpt-3's communication with diverse social groups on contentious topics. *Scientific Reports*, 14(1):1561.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024a. Don't just say "i don't know"! self-aligning large language models for responding to unknown questions with explanations. *Preprint*, arXiv:2402.15062.

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024b. Gotcha! don't trick me with unanswerable questions! self-aligning large language models for responding to unknown questions. *arXiv preprint arXiv:2402.15062*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. 2024. Mathematical capabilities of chatgpt. *Advances in neural information processing systems*, 36.

Yangyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan Kankanhalli. 2024. Unk-vqa: A dataset and a probe into the abstention ability of multi-modal large models. *Preprint*, arXiv:2310.10942.

Walid Hariri. 2023. Unlocking the potential of chatgpt: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. *arXiv preprint arXiv:2304.02017*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Jingshan Huang and Ming Tan. 2023. The role of chatgpt in scientific communication: writing better scientific review articles. *American journal of cancer research*, 13(4):1148.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Katikapalli Subramanyam Kalyan. 2023. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, page 100048.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Zihao Li. 2023. The dark side of chatgpt: Legal and ethical challenges from stochastic parrots and hallucination. *arXiv preprint arXiv:2304.14347*.

Ryan Lingo. 2023. The role of chatgpt in democratizing data science: an exploration of ai-facilitated data analysis in telematics. *arXiv preprint arXiv:2308.02045*.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023a. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017.

Yixin Liu, Avi Singh, C. Daniel Freeman, John D. Co-Reyes, and Peter J. Liu. 2023b. Improving large language model fine-tuning for solving math problems. *Preprint*, arXiv:2310.10047.

Yixin Liu, Avi Singh, C Daniel Freeman, John D Co-Reyes, and Peter J Liu. 2023c. Improving large language model fine-tuning for solving math problems. *arXiv preprint arXiv:2310.10047*.

Phuoc Pham Van Long, Duc Anh Vu, Nhat M. Hoang, Xuan Long Do, and Anh Tuan Luu. 2024. Chatgpt as a math questioner? evaluating chatgpt on generating pre-university math questions. *Preprint*, arXiv:2312.01661.

Jingyuan Ma, Damai Dai, and Zhifang Sui. 2024. Large language models are unconscious of unreasonability in math problems. *arXiv preprint arXiv:2403.19346*.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2024a. Do llms know when to not answer? investigating abstention abilities of large language models. *arXiv preprint arXiv:2407.16221*.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2024b. Do llms know when to not answer? investigating abstention abilities of large language models. *Preprint*, arXiv:2407.16221.

OpenAI. 2023. Chatgpt (mar 14 version) [large language model].

OpenAI. 2024. Openai api documentation. https://platform.openai.com/docs/. Accessed: 2024-10-13.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Paulo Shakarian, Abhinav Koyyalamudi, Noel Ngu, and Lakshmivihari Mareedu. 2023. An independent evaluation of chatgpt on mathematical word problems (mwp). *arXiv preprint arXiv:2302.13814*.

Yuhong Sun, Zhangyue Yin, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Hui Zhao. 2024. Benchmarking hallucination in large language models based on unanswerable math word problem. *arXiv preprint arXiv:2403.03558*.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust llms: Aligning confidence with response quality. *arXiv preprint arXiv:2404.17287*.

Leyao Wang, Zhiyu Wan, Congning Ni, Qingyuan Song, Yang Li, Ellen Wright Clayton, Bradley A Malin, and Zhijun Yin. 2024a. A systematic review of chatgpt and other conversational large language models in healthcare. *medRxiv*.

Zhengxiang Wang, Jordan Kodner, and Owen Rambow. 2024b. Evaluating llms with multiple problems at once: A new paradigm for probing llm capabilities. *arXiv preprint arXiv:2406.10786*.

Yousef Wardat, Mohammad A Tashtoush, Rommel AlAli, and Adeeb M Jarrah. 2023. Chatgpt: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7):em2286.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by llms: A showcase of chatgpt in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 610–625.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Xin Xu, Tong Xiao, Zitong Chao, Zhenya Huang, Can Yang, and Yang Wang. 2024a. Can llms solve longer math word problems better? *arXiv preprint arXiv:2405.14804*.

Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Wenyi Zhao, Jie Tang, and Yuxiao Dong. 2024b. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline. *Preprint*, arXiv:2404.02893.

Zihao Zhou, Qiufeng Wang, Mingyu Jin, Jie Yao, Jianan Ye, Wei Liu, Wei Wang, Xiaowei Huang, and Kaizhu Huang. 2024. Mathattack: Attacking large language models towards math solving ability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19750–19758.