FAULTY SCIENCE QA: CAN CHATGPT DETECT LOGICAL INCONSISTENCIES IN SCIENCE QUESTIONS *

Sinjoy Saha
The Pennsylvania State University

ABSTRACT

Consider the question: "Solve the equation $|y - 6| + 2y = 9$ for y, given that $y > 6$ and $y < -3$." This problem is inherently flawed: y cannot simultaneously satisfy $y > 6$ and $y < -3$. However, this raises an important question: Can current LLMs reliably detect such logical inconsistencies in faulty science and math problems or do they merely apply mathematical and scientific formulas mechanically without understanding? Although LLMs may detect a few simple, direct inconsistencies and faults in questions and refuse to answer them, in this work, we show that for a vast majority of faulty math, physics and chemistry questions, they mechanically apply the formula to solve similar types of questions. We create a diverse dataset of faulty math, physics and chemistry problems by introducing logical inconsistencies in the original problems. These problems are from various topics in the three subjects including concepts like algebra, geometry, force, work and energy, motion, atomic model, stoichiometry, ideal gas law, and thermodynamics. They also include a range of difficulty levels and various sources of faultiness, including common-sense violations, ambiguous statements, and mathematical contradictions. We primarily evaluate ChatGPT (4o and 4o-mini) across dimensions. Through extensive experimentation and analysis, we find that ChatGPT cannot be relied on to detect faults and inconsistencies in the questions itself, since it lacks the reasoning skills necessary to operate as logical thinkers. This highlights significant limitations in their ability to reason about mathematical problems beyond rote computation.

***K*eywords** LLM · faulty question · inconsistency · logic · reasoning · ChatGPT

1 Introduction

Text generation is experiencing a transformative shift with the emergence of advanced Large Language Models (LLMs) such as Generative Pre-trained Transformer (GPT-4) [1], LLaMA-3 [2], PaLM [3], and T5 [4]. These models, characterized by their massive parameter scales, excel at producing text that closely resembles human language and demonstrate exceptional performance across a wide range of tasks. By leveraging zero-shot and few-shot learning, they effectively generalize to new tasks with minimal reliance on task-specific training, marking a significant leap in AI capabilities. They have also been shown to be effective at solving math and science questions.

However, as shown in Figure **??**, a faulty math question, which from a human perspective can be easily detected as having a mathematical contradiction, leads ChatGPT to mechanically solve the question, completely ignoring the contradiction that has been introduced in the given question.

Inspired by [5], in this paper, we study the (in)ability of LLMs to answer faulty questions while either completely ignoring the fallacies or answering a modified question. Specifically, we create a dataset of math, physics and chemistry questions, and then modify the questions by introducing various assumptions or conditions that are inconsistent with or directly violate one or more mathematical or scientific laws. We find that ChatGPT cannot be relied on to detect faults and inconsistencies in the questions itself, since it lacks the reasoning skills necessary to operate as logical thinkers. This highlights significant limitations in their ability to reason about mathematical problems beyond rote computation.

*Submission for the final project for CSE 584 - Machine Learning: Tools and Algorithms.

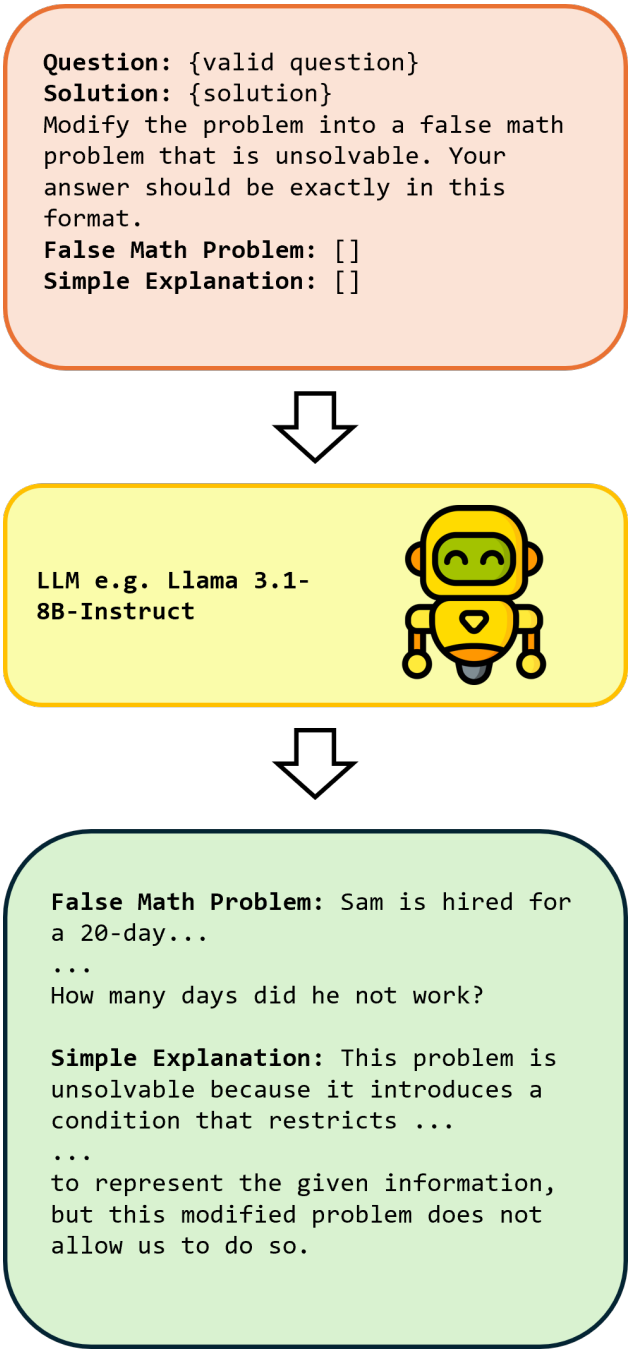

Figure 1: Prompt for automated Math dataset creation using Llama 3.1-8B-Instruct model.

2 Related Work

Mathematical and scientific problem-solving has long been a topic of AI research. There have been many historical works in this domain. One of the earliest attempts was by Bobrow et. al. [6] in building an expert system with rules and a parser for converting word problems into algebraic equations for solving using a computer. With the emergence of deep learning, models began to utilize neural networks. Prior works have been done to tackle the increasing complexity of mathematical problems.

More recently, general-purpose large language models (LLMs) like GPT-4, Gemini [7], and Llama 3 (Meta, 2024) and Claude 3 (Anthropic, 2024), have been developed, pushing the boundaries of AI's problem-solving capabilities.

Benchmarks such as GSM8K [8] and MATH [9] have become key tools for evaluating LLMsâĂŹ mathematical reasoning abilities. Furthermore, datasets like ScienceQA [10] for scientic question answering has also been developed. Additionally, specialized LLMs trained on mathematical datasets, including MathVerse [11], and DeepSeekMath [12] have emerged to address domain-specific challenges.

To improve LLM performance on these benchmarks, several techniques have been introduced:

- Prompt Engineering: Techniques like zero-shot and few-shot learning [13][11] refine prompts to guide models toward accurate and contextually appropriate outputs.
- Chain-of-Thought (CoT) Prompting: This method [14] structures problems into sequential intermediate steps to enhance logical reasoning.
- Deeply Understanding Problems (DUP): DUP [15] focuses on ensuring a thorough comprehension of problem intricacies before attempting solutions.
- Program-Aided Language Modeling: Leveraging LLMs' code completion capabilities, this approach improves reasoning by integrating programming-based methods.
- Agent-Aided Approaches: Multi-agent systems collaboratively enhance mathematical reasoning and problem-solving.

Despite these advancements, most efforts have centred on solving problems which are valid, whether mathematical or scientific, with limited exploration of LLMsâĂŹ logical reasoning capabilities in more complex concepts.

In this paper, we study the effects of asking invalid or faulty science questions to one the leading close-source LLMs, ChatGPT [1], and analyse it performance.

3 Dataset

We create a diverse dataset of 113 faulty math, physics and chemistry problems by introducing logical inconsistencies in the original problems. These problems are from various topics in the three subjects:

1. Multiple mathematical concepts like algebra, geometry, number theory, functions, and linear equations.
2. Multiple physics concepts like forces, work and energy, motion, and simple harmonic motion.
3. Multiple chemistry concepts like atomic model, stoichiometry, gas law, and thermodynamics.
4. A range of difficulty levels.
5. Various sources of faultiness, including common-sense violations, ambiguous statements, and mathematical contradictions.

A detailed workflow of the dataset creation for each of the three subjects - math, physics, and chemistry are the subsequent sections.

3.1 Mathematics

We built this dataset using the MATH [9] dataset as original reference for valid math questions. The dataset is downloaded from Hugging Face [2]. We use $\sim$5000 questions and prompt Llama 3.1 8B Instruct-tuned model to create a faulty question from the original question by introducing a mathematical inconsistency. However, due to resource constraints, we consider only $\sim$26 questions and give them to ChatGPT for answering and further analysis. The entire workflow and prompts for generation using Llama and final parsing are given in Figure 1.

Next, we give these questions to ChatGPT and

3.2 Physics

The physics questions are taken from NCERT[3] high school books and then given to Claude Sonnet 3 for generating invalid questions.

The prompt use for the same is as follows:

[2]https://huggingface.co/datasets/hendrycks/competition_math

[3]https://ncert.nic.in/

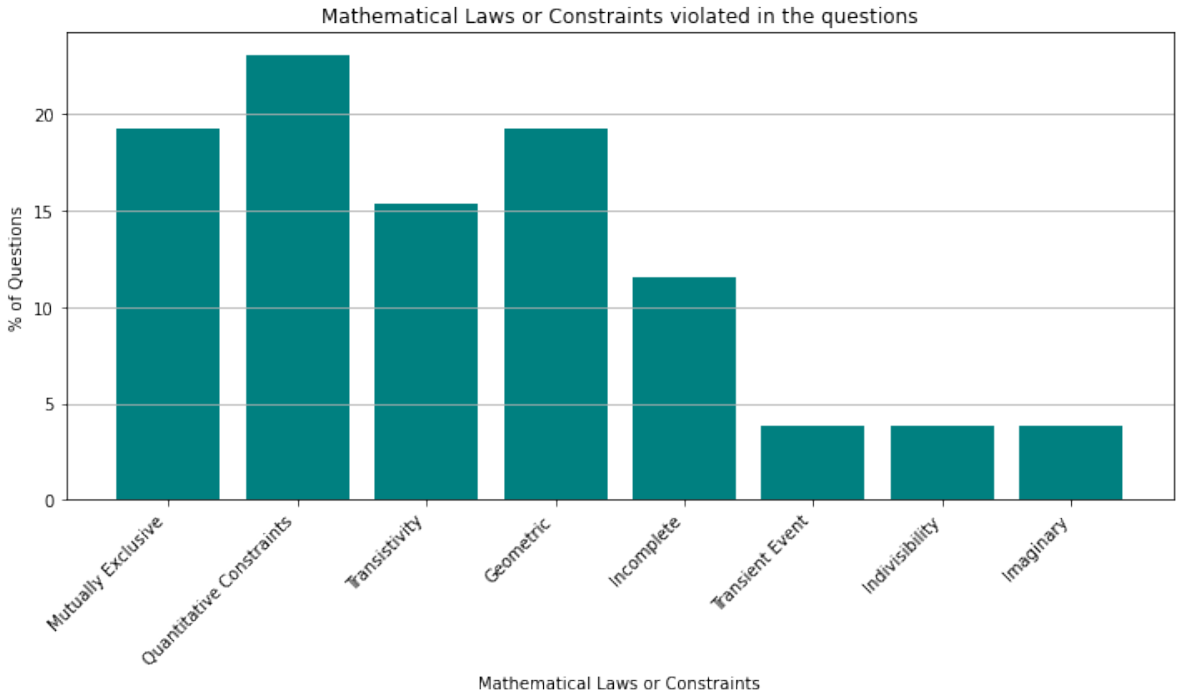

Figure 2: Mathematical laws or constraints violated in the given questions.

```
You are an expert in science and math.  Modify a given list of questions to create
a subset of unsolvable questions.  These unsolvable questions must:  Be subtle
modifications of the originals, appearing solvable at first glance.  Contain violations
of established principles (e.g., logical inconsistencies, physical impossibilities).
Avoid being obviously unsolvable without careful analysis.  [Questions]
```

Once Claude returns these questions, we feed these in to ChatGPT for getting the answer.

3.3 Chemistry

The chemistry questions are taken from NCERT[4] high school books and then given to Claude Sonnet 3 for generating invalid questions.

The prompt use for the same is as follows:

```
You are an expert in science and math.  Modify a given list of questions to create
a subset of unsolvable questions.  These unsolvable questions must:  Be subtle
modifications of the originals, appearing solvable at first glance.  Contain violations
of established principles (e.g., logical inconsistencies, physical impossibilities).
Avoid being obviously unsolvable without careful analysis.  [Questions]
```

Once Claude returns these questions, we feed these into ChatGPT for getting the answer.

4 Results

In this section, we summarize the results of two studies. First, we study the effect of including text generated by the LLMs for mathematical prompts in the dataset. Second, we study the effect of sequence length on classification performance.

[4]https://ncert.nic.in/

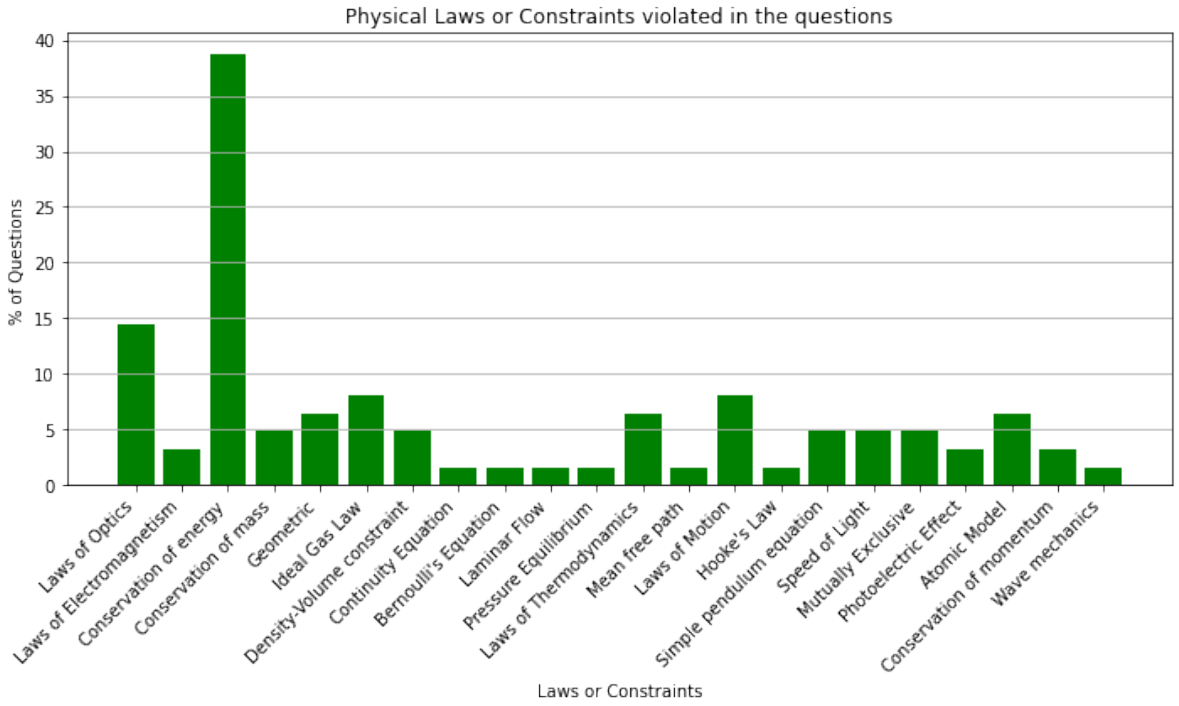

Figure 3: Physical laws or constraints violated in the given questions.

4.1 Verbosity Analysis

In this section, we study the verbosity of ChatGPT with respect to the domain of the invalid question.

Verbosity analysis involves evaluating the length and conciseness of textual outputs or communication to determine whether they convey information effectively. It focuses on identifying unnecessary or redundant elements that might obscure the main message or reduce clarity. This analysis is essential in fields like natural language processing, education, and business communication, where excessive verbosity can hinder comprehension or efficiency. By quantifying verbosity and examining its impact, verbosity analysis helps improve communication quality, ensuring that information is concise, relevant, and easy to understand.

We see that the math questions have a very normal distribution with respect to verbosity. This may be due to the over-training of these models on math dataset.

We see that the physics question answers have a somewhat similar normal distribution with respect to verbosity. This may be due to some training of these models on physics dataset.

We see that the median verbosity of chemistry questions is around 300 words. However, there are few questions which are much more verbose at 500 words than the other two disciplines.

5 Conclusion

In this paper, we study the (in)ability of LLMs to answer faulty questions while either completely ignoring the fallacies or answering a modified question. Specifically, we create a dataset of math, physics and chemistry questions, and then modify the questions by introducing various assumptions or conditions that are inconsistent with or directly violate one or more mathematical or scientific laws. We find that ChatGPT cannot be relied on to detect faults and inconsistencies in the questions itself, since it lacks the reasoning skills necessary to operate as logical thinkers. This highlights significant limitations in their ability to reason about mathematical problems beyond rote computation.

References

[1] OpenAI. Gpt-4 technical report, 2024.

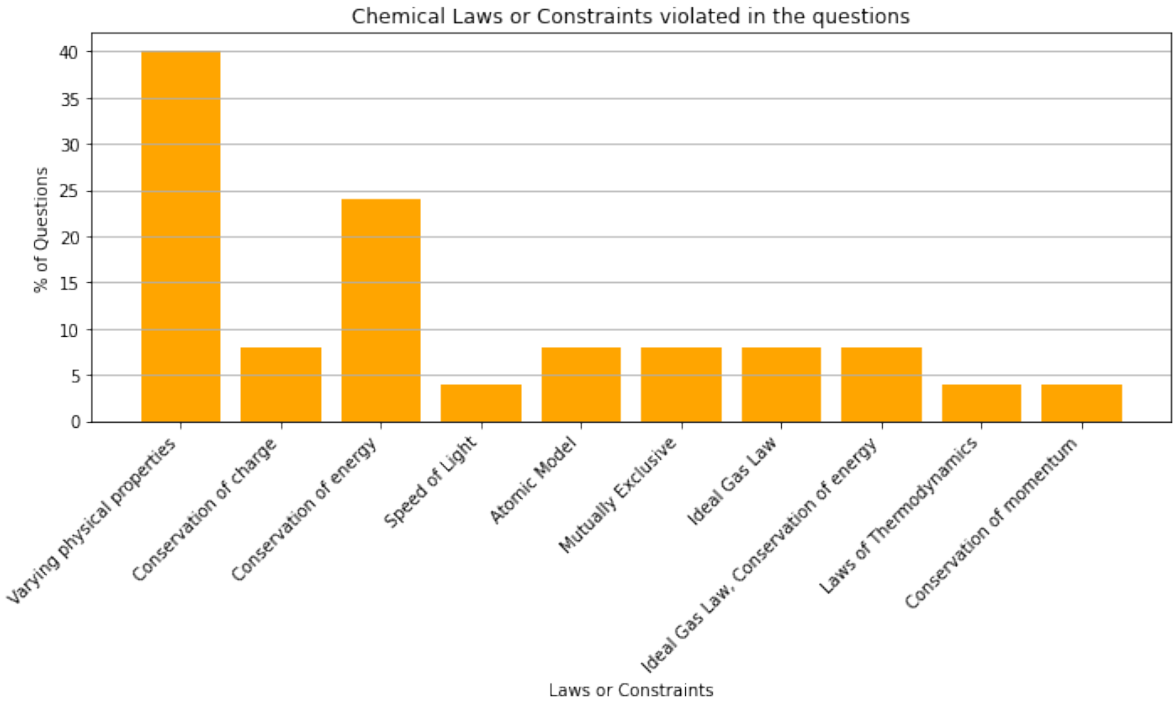

Figure 4: Chemical laws or constraints violated in the given questions.

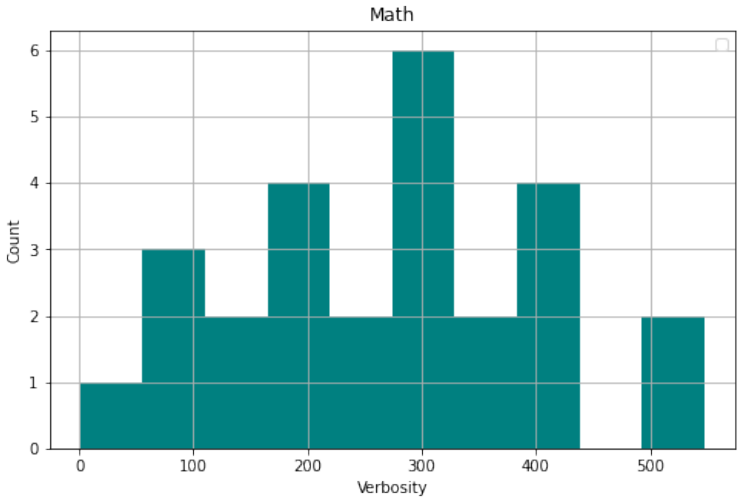

Figure 5: Verbosity analysis for Math questions.

[2] Abhimanyu Dubey et al. The llama 3 herd of models, 2024.

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[5] AM Rahman, Junyi Ye, Wei Yao, Wenpeng Yin, and Guiling Wang. From blind solvers to logical thinkers: Benchmarking llms' logical integrity on faulty mathematical problems. *arXiv preprint arXiv:2410.18921*, 2024.

[6] Daniel G Bobrow. A question-answering system for high school algebra word problems. In *Proceedings of the October 27-29, 1964, fall joint computer conference, part I*, pages 591–614, 1964.

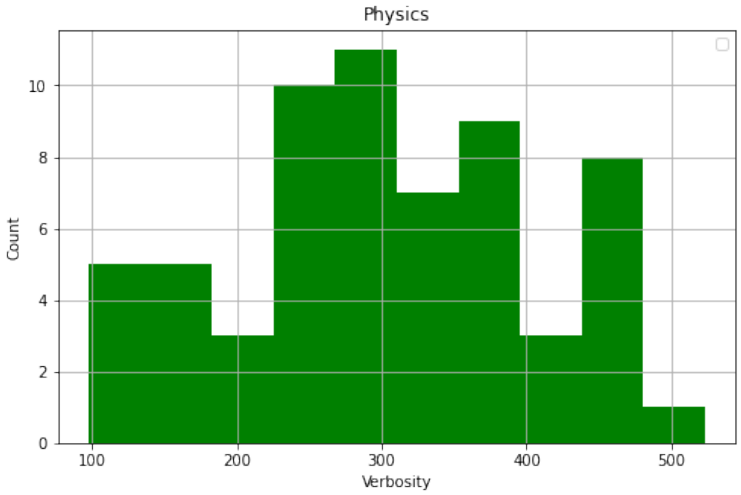

Figure 6: Verbosity analysis for Physics questions.

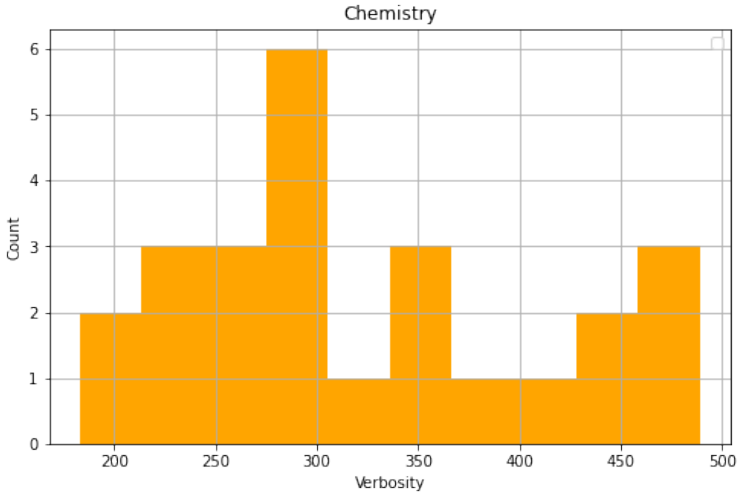

Figure 7: Verbosity analysis for Chemistry questions.

[7] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

[10] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[11] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

[12] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

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

[15] Qihuang Zhong, Kang Wang, Ziyang Xu, Juhua Liu, Liang Ding, Bo Du, and Dacheng Tao. Achieving> 97% on gsm8k: Deeply understanding the problems makes llms perfect reasoners. *arXiv preprint arXiv:2404.14963*, 2024.