Analysis of LLM Response Patterns to Faulty Scientific Questions

-Soham Barman

This study investigates the performance of Large Language Models (LLMs) when confronted with faulty scientific problems in physics and chemistry domains. By systematically generating intentionally flawed questions from academic textbooks, the research explores how LLMs handle conceptually impossible or logically inconsistent scientific scenarios. Through detailed linguistic and sentiment analysis, the study reveals significant limitations in LLMs' reasoning capabilities, including low accuracy in identifying logical inconsistencies, variable confidence levels, and an inability to consistently recognize fundamental scientific errors. The findings highlight critical gaps in current AI reasoning, demonstrating that despite advanced problem-solving abilities, LLMs struggle to detect subtle conceptual contradictions across different scientific domains. The research underscores the urgent need for developing more robust reasoning mechanisms in artificial intelligence systems.

Introduction

In recent years, Large Language Models (LLMs) have demonstrated significant limitations in handling scientific problems, particularly in physics and chemistry domains. This presents an intriguing opportunity to explore how these models handle fundamentally flawed or unsolvable problems. While LLMs can achieve up to 62.5% accuracy on well-specified physics problems, their performance drastically drops to 8.3% when dealing with under-specified or ambiguous problems [1].

The phenomenon becomes particularly interesting when we consider that these models often generate confident-sounding, well-written responses even when presented with faulty premises [2]. This behaviour stems from their fundamental design - they are trained to produce helpful and convincing responses without guaranteed accuracy. In the domain of chemistry and physics, LLMs exhibit specific failure modes, including errors in algebraic manipulation, difficulty in grounding abstract concepts into mathematical equations, and failure in retrieving relevant domain-specific concepts [3].

This research explores a novel approach: deliberately constructing faulty physics and chemistry questions that exploit these known limitations. The significance of this investigation is amplified by recent findings showing that users often overlook mistakes in LLM responses, accepting incorrect answers 39% of the time. This human tendency to anthropomorphize and trust LLM outputs [2] creates a unique opportunity to study how these models handle intentionally malformed problems, potentially revealing deeper insights into their reasoning processes and limitations.

Discussion of Related Works:

Liu et al. [4] critically investigate the uncertainty reflected by LLMs when questions beyond their parametric knowledge are thrown at them. In this paper, the authors established a new benchmark called UnknownBench, comprising three carefully designed tasks, and reviewed how well LLMs can strike a balance between transparency and utility in those cases where the questions cannot be answered. The approach of the study is really interesting, consisting of three different tasks: NEC contains 2,078 questions with artificially non-existent concepts and 2,072 control questions with real concepts; FalseQA includes 4,730 questions whose premise is all false; and RefuNQ, a companion to the NaturalQuestions dataset, includes carefully constructed unanswerable queries. The authors employ an end-to-end evaluation framework that uses lexical keyword matching to assess refusal, measures helpfulness using answer accuracy on answerable questions, and uses one unified confidence elicitation approach for all LLMs.

These findings demonstrate serious defects in state-of-the-art LLMs in terms of their expressive capability for uncertainty. GPT-4, the most advanced model that participates in this work, refuses about 35% of unanswerable questions, whereas open-source models such as Llama-2, Vicuna, and Mistral achieve an average refusal rate of only 28.3%. This seems to present an investigation that offers persuasive evidence in proving that SFT and RLHF are gradually improving refusal capabilities-a fact demonstrated in the chat variants of Llama-2, setting new records in front of their founding models.

One of the most interesting aspects of the research is the subtlety brought about in confidence expression. The authors reveal a large gap between articulated confidence and perceived confidence in the answers of LLMs. Interestingly, models often appear confident when they are actually deeply uncertain, with the GPT-4 evaluation showing inconsistent correlations between perceived confidence and true accuracy. Proprietary large language models showed even more developed behaviors, with a positive relationship between accuracy and confidence, and a negative one between refusal rates and confidence-an observed pattern distinctly missing in open-source models.

Complementary to that, Williams and Huckle [5] give a very strong discussion of the weaknesses in the abilities of LLMs by introducing a very well-thought-out 30-question Linguistic Benchmark. This innovative approach is targeting performance on tasks that humans find simple and straightforward, thus laying open crucial gaps in current artificial intelligence. These results were quite striking, reaching a maximum of 38% for even the most recent models like GPT-4 Turbo over domains such as logical reasoning, spatial intelligence, and linguistic and mathematical reasoning.

The proposed assessment framework probes a manifold array of state-of-the-art LLMs of leading tech companies with a comprehensive view toward strengths and weaknesses; one central conclusion was that LLMs tend to fail basic tasks in challenges that test reasoning skills, indicating that they often retreat to offering solutions that fit their training data rather than considering the peculiarities of novel questions. Therefore, this challenges the dominant discourse on the superiority of LLMs and, hence, calls for more advanced mechanisms in reasoning.

Most encouraging in the present research has been a great improvement in performance due to the sophisticated skill of prompt engineering. The authors allowed models to first ask for clarification on a question before providing a specific answer and found that on average, all models tested gained a relative performance improvement of 40.7%. This was observed more in models like Llama 3 70B and Mistral 8x22B that are open-source, showing ample scope for future developments in the realm of AI reasoning.

The work of Rahman et al. [6] further extends this investigation into the reasoning capabilities of LLMs through their FAULTYMATH dataset, focused on the identification of logically inconsistent mathematical problems by models. The study presents a provocative argument that current LLMs function more as "blind solvers" than "logical thinkers," with even the most advanced model, Gemini 1.5 Pro, correctly identifying only 33.33% of faulty problems.

The authors of this work indicate one critical vulnerability in the reasoning of LLMs: high susceptibility to confirmational bias. Whereas model performance did much better when models were given correct hints, they were very susceptible to being misguided by hints. This finding underlines the current limitations in AI logical reasoning; this would mean very good performance on standard mathematical datasets does not imply the ability to detect basic logical inconsistencies.

Significantly, the research demonstrated subtle differences in mathematical reasoning as it pertains to various problem categories. Algebraic problems exhibited the greatest accuracy among the models examined, whereas geometry posed the most considerable difficulties, characterized by an average accuracy of merely 23%. This discrepancy indicates that specific domains of mathematical reasoning continue to be especially problematic for contemporary large language model architectures.

Another interesting perspective of this research is the comparison between the human performance and the AI performance. Human performances averaged 21.33% for finding the false mathematical problems, while the average performance of all LLMs stood at 19.72%. This modest difference underlines both the development of AI capability and intrinsic complexity in logical reasoning tasks.

What these detailed examinations amount to in their totality is indicating the dire need for further innovation in the development of LLMs. They variously indicate the reinforcement of basic reasoning skills, subtle mechanisms for the expression of uncertainty, and stronger means for the conduction of logical reasoning. The paper reminds one that though LLMs have achieved great milestones, much remains to be overcome on the road to genuinely intelligent and context-sensitive artificial intelligence.

Another paper which came close to discussing this issue was by Li et al. [7] put up a very logical reasoning-based argument for the need to focus on logical fallacy understanding in overcoming LLM limitations. The proposed work envisages a comprehensive framework of looking at fallacies in three cognitive dimensions, viz., WHAT-identification, WHY-deductive reasoning, and HOW-correction. The authors extend this by providing a new evaluation and improvement method for the LLM reasoning capability by constructing the Logical Fallacy Understanding Dataset, LFUD, which consists of twelve logical fallacy types. As demonstrated in this work, fine-tuning with LFUD significantly improves logical reasoning, with the larger models such as GPT-4 achieving accuracy rates above 85% in most tasks. One of the key findings is that performance on logical reasoning tasks improves with increased training data. In this way, the researchers provide theoretical and empirical tools for developing more sophisticated artificial intelligence reasoning systems and underline how important logical fallacies are for further development in AI.

Dataset Curation

Choice of Valid Questions:

In my research, I tried to develop a comprehensive dataset of invalid scientific problems from textbook sources across multiple academic levels: Selina ICSE Concise Physics for Class 9, NCERT Physics Classes 11 and 12, and NCERT Chemistry Classes 11 and 12. My approach to generating invalid problems was born out of necessity and computational constraints and thus necessarily called for a multi-stage process with careful orchestration.

First of all, the use of the Anthropic Claude Sonnet 3.5 as our main LLM was the first stage of problem generation in my contribution. We were to manually input questions from some selected textbooks and ask the model to convert each valid problem into an essentially unsolvable one, including a full explanation of why such a problem would be inherently flawed. From the approximately 400 generated problems that I obtained, I filtered out 237 of sufficient difficulty and potentially invalid.

In the last step for producing responses to this defective question, I had to systematically feed these 237 generated problems to ChatGPT 4o-mini and analyze its response mechanisms in a bid to make out which ones will provide consistently wrong or nonsensical solutions. I found through this process that 152 problems elicited fundamentally flawed or inappropriate responses; thus, these became the major elements of our experimental data.

Problem generation in this manner was intrinsically iterative and time-consuming. The generated problems underwent several layers of filtering: generation via Claude Sonnet 3.5 with the goal of creating plausible versions of unsolvable versions of normal textbook problems; my personal manual review to very carefully exclude obviously trivial or completely nonsensical examples; and probing via ChatGPT 4o-mini upon the nature and depth of computational failures likely to be evoked in trying to solve said problems.

I readily acknowledge the limitations of our current methodology. my research was constrained by limited API access, which necessitated a time-intensive manual approach to dataset creation. In my opinion, the ideal way of generating challenging problem sets, I propose a comprehensive approach to creating intentionally flawed mathematical or scientific problems. The methodology begins with comprehensive source material collection, specifically compiling problem sets from diverse textbooks across multiple academic domains. This first step is essential to wide and representative sampling of all possible source problems.

My contribution's essential novelty is to utilize large language models' capabilities in systematically modifying those problems into ones that are posed in a way that these problems are challenging or not solvable at all. For instance, the LLM will generate intentionally faulty problem statements by feeding the original problems and their correct solutions. This approach leverages the capabilities of the LLM regarding the manipulation and reinterpretation of complex problem formulations with an apparent academic credibility.

In this regard, I suggest multistage verification: The first stage of verification is through a human expert review, whereby researchers check the LLM-generated problems for academic soundness and for flaws that are intentionally introduced. I would further validate such problematic questions through another LLM system, which would have the task of identifying and articulating for each problem the unsolvability or logical inconsistencies. This secondary computational analysis can give nuanced insights into the structural challenges of the generated problems. The final verification would then involve domain experts who can finally ascertain the suitability of the problems either for educational or research purposes. This rigorous validation makes sure that the generated problem set meets high academic standards and maintains the intended goal of creating challenging, thought-provoking mathematical or scientific exercises.

Experiments and Dataset Analysis

The research questions that are explore in the report are: **What are the main factors that affect the confidence and technical density of a response? How do misconceptions arise in specialized domains, and can they be mitigated? What roles do sentiment and linguistic markers play in shaping perceptions of these responses?** I hope my analysis will bring to light the hidden pattern of these questions and give some insight into the output of models across different domains.

In my investigation, I analyzed the subtlety of feeling, polarity, and subjectivity expressed with the help of specialized terminology, the frequency of uncertainty markers, the impact on the score of confidence. I visualize the trends in both physics and chemistry to point out the difference in the model responses concerning different types of questions. Another great motivation that arises in my research is **the detection of misconceptions through linguistic analysis, where I show just how far answers may be off from recognized domain-specific standards**. These observations raise very serious questions about the reliability and domain-specific calibration of LLMs.

There is a lot of nuances involved in the analysis of performance of LLMs on physically impossible scenarios with respect to the way AI reasons over the input providing it. One of the striking findings I found involved comparing the confidence levels of questions and responses. For example, I have encountered a question confidence of 70.000, which, in turn, was gave a response confidence of 27.850 to yield a very high confidence differential of 42.150. In this case, **the large language model is quantifiably uncertain in situations** that are anathema to fundamental physical intuition.

Linguistic and sentiment analyses show nuanced patterns in their communication. I observed that the polarity scores concentrated around neutral values between -0.061 and 0.160, **showing measured speech against situations that might be in conflict**. The measures of subjectivity that I recorded were between 0.270 and 0.695, which further **indicates more interpretive language** by the models. The length reached a maximum of 266 words-maybe a hint **regarding the implication of complexity in describing scenarios that are physically impossible**.

However, the uncertainty density metrics were even more interesting in the range from 0.003759 to 0.011696. Analyzing, I found indeed that **higher uncertainty densities correspond to more complex physical contradictions** in their meaning, thus again hinting at a principled approach toward solving a conceptual challenge. Present, this observation is especially sharp for problem-solving contexts that incorporate basic physical impossibilities, such as the spread of sound in a vacuum or the opposite acoustic effect.

Continuing with this, I started to notice disciplinary differences in some properties of questions and responses. As an example, Physics questions are wordier: their average is 76.4 words, whereas for Chemistry it is more succinct-57.0 words. Curiously, I **noticed this relationship flips in responses: the average word count for Chemistry explanations is 419.4 words as opposed to Physics responses, averaging 368.5 words, which may suggest that chemical concepts are more complexly elaborated despite simpler initial formulations**.

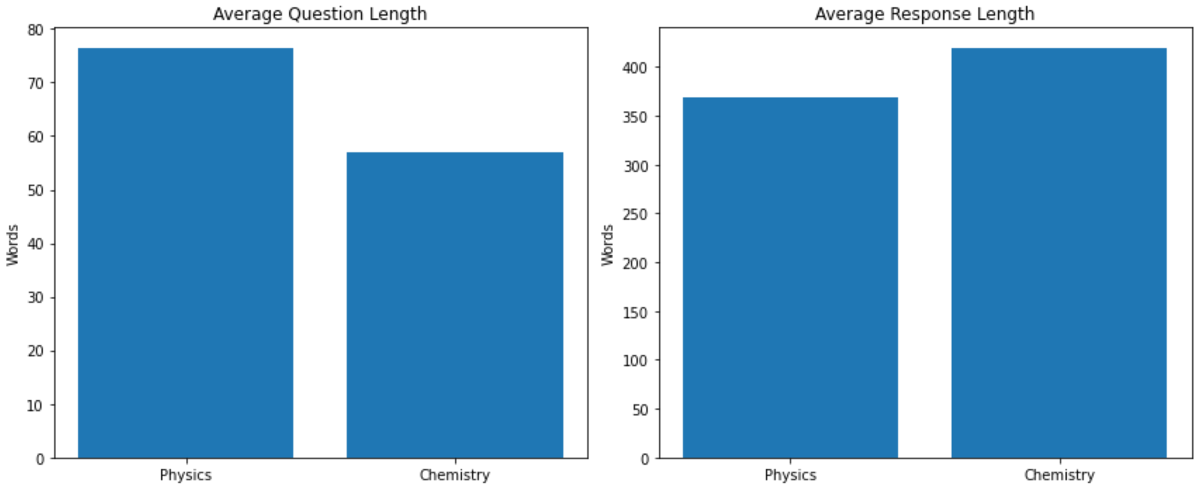
I have performed an analysis of misconceptions whereby I found typical patterns of errors in different subjects. In **Physics, main errors are related to wave mechanics, conservation of energy, and manipulations of physical constants**. I found typical categorization of errors in

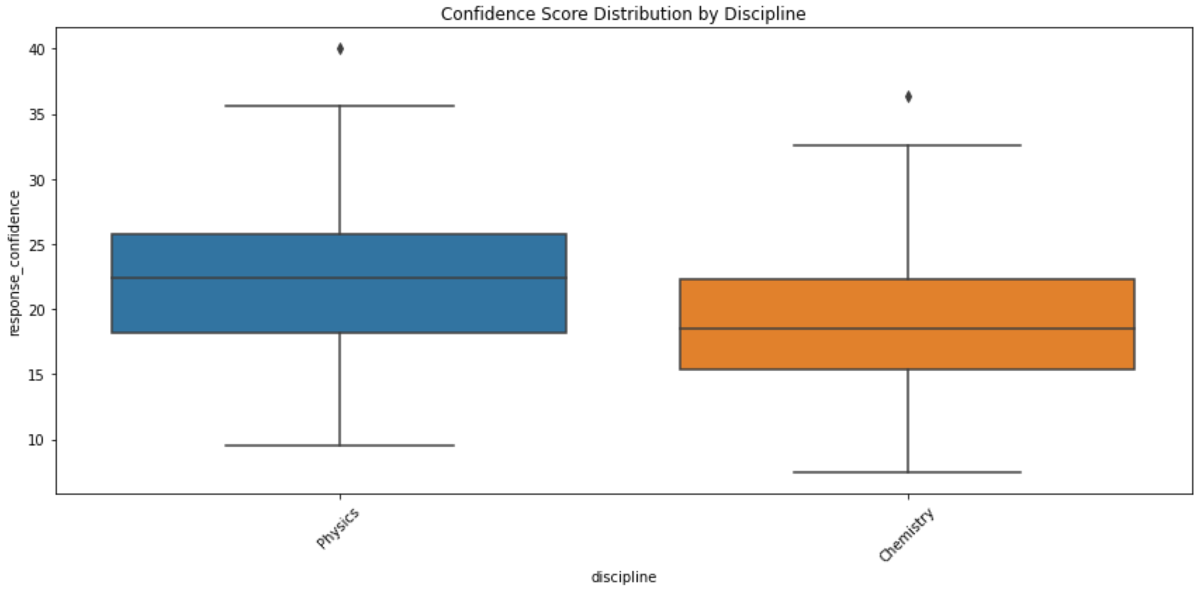
Chemistry: misconceptions of **pressure-temperature relationships, logical impossibility in the percentage calculations, and fundamental violations in rate changes**. These are the domains where the LLMs fail to capture that something inherently wrong is with the question when subtle changes are done to these areas of the question.

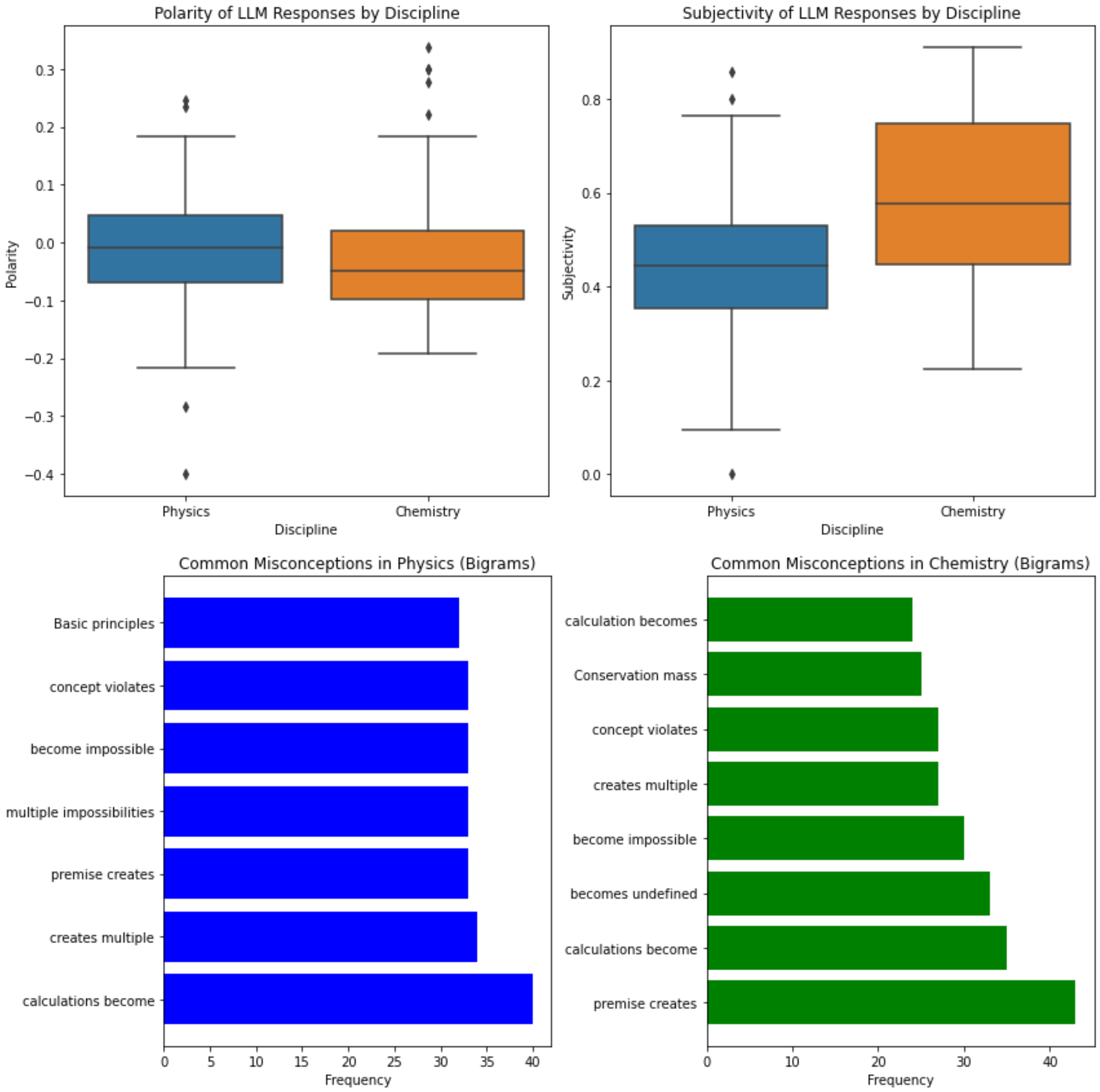
Most remarkably, I found that **questions with higher technical density tend to receive responses that contain relatively low levels of technical complexity**. I also noticed that, interestingly, the technical density of responses remained constant and rarely grew beyond the level of technical sophistication encountered in the original question. This led me to surmise a cautious, adaptive strategy followed by LLMs while answering complicated scientific scenarios challenging the basic physical and chemical principles.

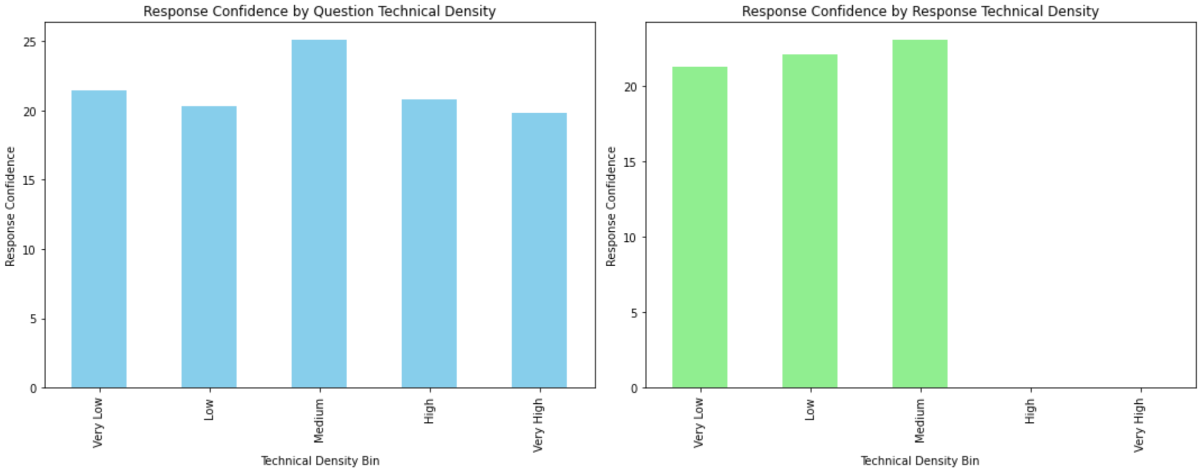
The current research enterprise informatively enlightens on the reasoning processes of artificial intelligence: while at work was advanced problem-solving ability, there was also deep-seated limitation to recognize simple conceptual inconsistencies. In my judgment, such findings have a number of critical implications for AI education in the development of more robust scientific reasoning schema.

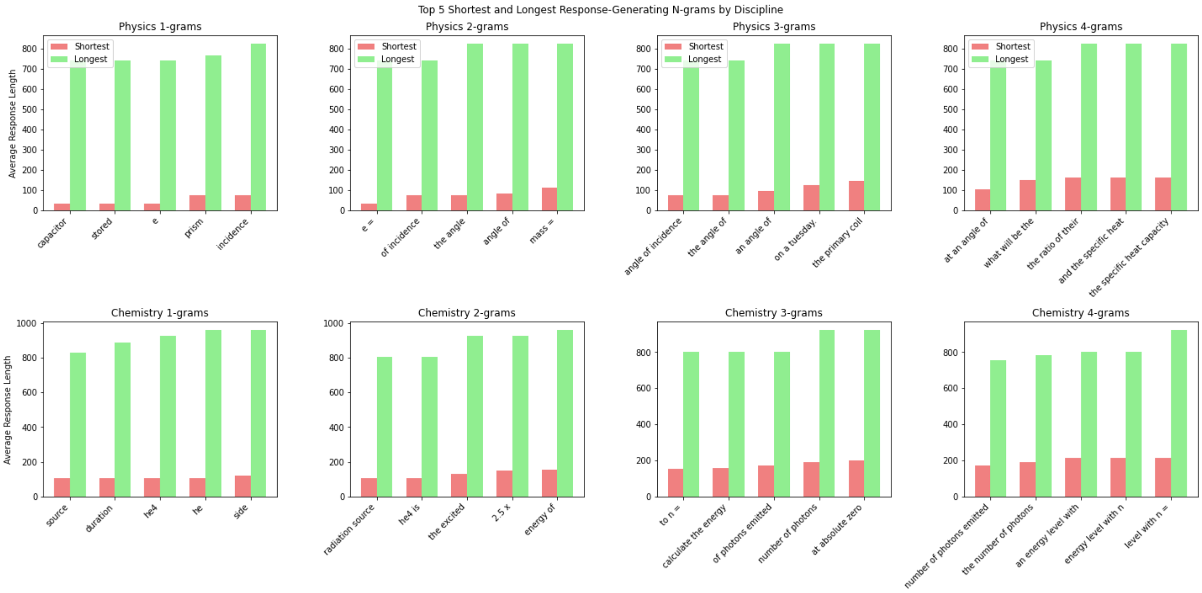
Here are a few figures below to better illustrate what was described above.






Polarity of LLM Responses by Discipline
Polarity
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
Physics
Chemistry
Discipline
Subjectivity of LLM Responses by Discipline
Subjectivity
0.8
0.6
0.4
0.2
0.0
Physics
Chemistry
Discipline
Common Misconceptions in Physics (Bigrams)
Basic principles
concept violates
become impossible
multiple impossibilities
premise creates
creates multiple
calculations become
0
5
10
15
20
25
30
35
40
Frequency
Common Misconceptions in Chemistry (Bigrams)
calculation becomes
Conservation mass
concept violates
creates multiple
become impossible
becomes undefined
calculations become
premise creates
0
10
20
30
40
Frequency





Limitations and Future Work

The present research faced a number of significant limitations that inherently **restricted the scope of our investigation**. Our process for creating the dataset was **mainly manual and resource-intensive**, based mostly on limited API access and very time-consuming verification processes. In addition, the methodology was mainly constrained to physics and chemistry textbooks from specific academic levels, which naturally restricts the generalizability of our findings.

Future research directions open many possibilities for extension and fine-tuning of our approach. We suggest an extended dataset creation that would involve developing a parallel control dataset of well-posed questions corresponding to each of the ill-formed ones. **This would allow a systematic comparative study between LLM responses to faulty and control questions, which may lead to finer details about AI reasoning processes**. Moreover, the scope of the research could be extended to include disciplines other than physics and chemistry, such as economics and biology, to give a broader view of LLM performance in different scientific domains.

One of the most promising avenues for future investigation is a **multi-model comparative analysis**. Given the limitation of resources, our experiments were limited to only ChatGPT 4omini for answering faulty question and claude 3.5 for building faulty questions. For a more comprehensive approach, **experiments would need to be run on a wide range of AI models, investigate response variation**, and provide a standardized comparison framework on how different LLMs perform when faced with scientifically impossible problems. This would allow a deeper understanding of how various AI systems handle conceptual inconsistencies.

We see significant room for improvement of our experimental protocol. **Successive iterations of our experiment could use more subtle hint-based testing to better see how subtle contextual cues affect the responses given by LLMs**. Designing a binary evaluation metric to directly assess models' abilities to recognize fundamentally flawed scientific problems would provide more straightforward insights into the models' ability to reason. To quantify AI reasoning, one could also work on a more fine-grained scoring of logical inconsistencies.

The current research was fundamentally limited by manual verification processes, limited computational resources, and restricted access to multiple LLM platforms. Addressing these limitations would involve the automation of the dataset generation and verification process, securing broader API access to multiple language models, and developing more sophisticated evaluation metrics for assessing logical reasoning. While our work provides insight into the limitations of LLM reasoning, at the same time, it opens several avenues for further exploration. These extensions will substantially advance the understanding of artificial intelligence reasoning in science and technical domains. This research emphasizes the critical need for ongoing development of more robust and logically

References

[1] Wang, K. D., Burkholder, E., Wieman, C., Salehi, S., & Haber, N. (2024). Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. Frontiers in Education, 8, 1330486. https://doi.org/10.3389/feduc.2023.1330486

[2] Large Language Models pose risk to science with false answers, says Oxford study | University of Oxford https://www.ox.ac.uk/news/2023-11-20-large-language-models-pose-risk-science-false-answers-says-oxford-study-0

[3] Arora, D., & Singh, H. G. (2023). Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. ArXiv. https://arxiv.org/abs/2305.15074

[4] Liu, G., Wang, X., Yuan, L., Chen, Y., & Peng, H. (2023). Examining LLMs' Uncertainty Expression Towards Questions Outside Parametric Knowledge. *ArXiv*. https://arxiv.org/abs/2311.09731

[5] Williams, S., & Huckle, J. (2024). Easy Problems That LLMs Get Wrong. *ArXiv*. https://arxiv.org/abs/2405.19616

[6] Rahman, A. M., Ye, J., Yao, W., Yin, W., & Wang, G. (2024). From Blind Solvers to Logical Thinkers: Benchmarking LLMs' Logical Integrity on Faulty Mathematical Problems. *ArXiv*. https://arxiv.org/abs/2410.18921

[7] Li, Y., Wang, D., Liang, J., Jiang, G., He, Q., Xiao, Y., & Yang, D. (2024). Reason from Fallacy: Enhancing Large Language Models' Logical Reasoning through Logical Fallacy Understanding. *ArXiv*. https://arxiv.org/abs/2404.04293