Name: Aishwarya Kulkarni
ID: 924430258

CSE 584: Final Project Report

Abstract –

The present study explores the weaknesses of LLMs in processing faulty science questions constructed to take advantage of their shortcomings in reasoning. For this purpose, a dataset of 166 faulty questions in 18 scientific disciplines was developed by converting valid science questions into faulty ones through the introduction of logical inconsistency, ambiguous phrasing, or terminological misuse. Each question that was faulted was analyzed by discipline and fault type, as well as the response of leading LLMs: ChatGPT, Gemini, Claude, and Perplexity. The results indicate that while LLMs excel in factual recall, they perform badly in tasks that require interpretative reasoning or ambiguity detection, especially in disciplines like Physics and Mathematics. Patterns of mistakes, overconfidence in wrong answers, and lack of logical reasoning point to large-scale flaws in the ways LLMs validate inputs.

Introduction –

LLMs, such as ChatGPT, Gemini, Claude, and others, have broken many barriers concerning natural language understanding. At the same time, none of these engines have passed beyond their Achilles heel of easily being misled into performing unwanted actions by some kinds of invalid inputs that could be ambiguous, flawed, or logically inconsistent. This project investigates these limitations through a systematic analysis of the responses of LLMs to a curated dataset of faulty science questions.

I leveraged the SciQ dataset from Hugging Face, one of the most popular benchmarks for evaluating AI systems on science-related queries. Starting with approximately 300 questions at from this dataset, using GPT-3, I changed the questions to introduce illogical bias, ambiguous phrasing, or a flawed premise. For instance, a chemistry question intentionally misuses terms, while a biology question contains conditions that cannot coexist. The curated dataset consists of 166 faulty questions representing various disciplines such as biology, chemistry, physics, and mathematics. Each question entry contains (1) the faulty question itself, (2) the reason it is faulty, (3) the LLM it was tested on, and (4) the faulty response generated by the model. This not only investigates the performance of different LLMs but also traces a pattern in the failure of models. This dataset has been used to investigate the limits of LLM understanding and inference in ambiguous or misleading situations. In this paper, I am going to present the way of dataset creation, the experimental setting to test LLMs, and the results based on the analysis.

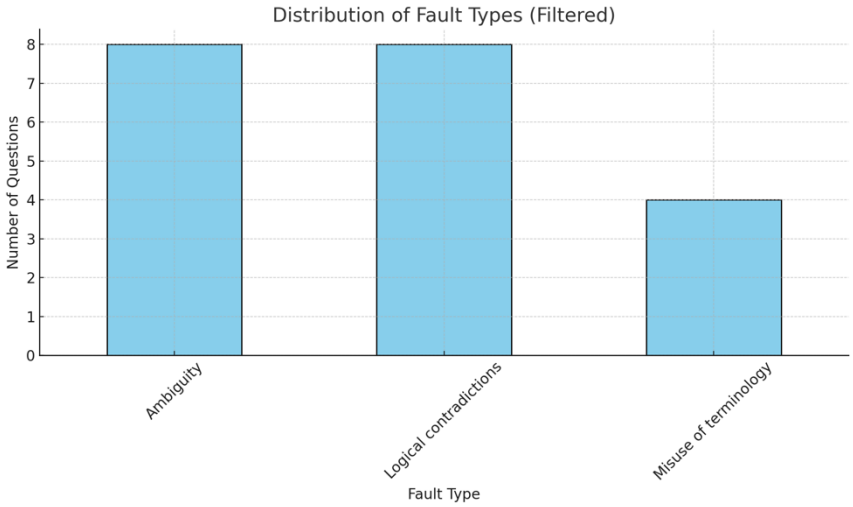
Research Questions –

1. **“What types of faulty science questions are most challenging for LLMs?”**

   I took reference of about 300 questions from Hugging Face’s SciQ dataset and induced some fault in the questions using prompting in GPT-3, be it a logical fault or a subtle vagueness in the questions, so that other existing LLMs would answer them incorrectly, without recognizing the fault. Then, I tested 4 AI models – Gemini, Claude Llama, and Perplexity using these questions and logged questions which could fool these AI tools. I built a dataset of such

incorrectly answered faulty questions by the LLMs. Examination of this dataset indicated the types of questions LLMs fared worst on were those relating to:

- **Logical Contradictions:** Questions that have internal inconsistencies, such as "A vehicle accelerates to a speed of -10 m/s," are often responded to as if valid. LLMs do not have mechanisms to ensure inputs are logically consistent.
- **Ambiguous Prompts:** Open-ended or vaguely phrased questions, such as "What contributes most to nitrogen fixation?", often yielded wrong answers. Instead of pointing out the ambiguity, LLMs would make unjustified assumptions.
- **Subtle Misuse of Terminology:** Questions with errors in scientific terms, such as confusing "sublimation" for "evaporation", were seldom detected as faulty. LLMs treated them as valid and responded with plausible but wrong explanations.



These findings suggest that LLMs are more challenged by interpretative reasoning and fault detection than by straightforward factual recall.

2. "How do different LLMs perform across disciplines?"

In the dataset I built, there are about 18 different disciplines. The ambiguous scientific questions across several disciplines were found to yield very different performances for Gemini, Claude, Llama, and Perplexity.

1. Model-Specific Observations

- Gemini: Strong factual recall, especially in biology and chemistry. For example, it gave correct answers to direct questions with embedded subtle ambiguities. Difficulty detecting unclear premises. Ambiguous prompts like "What is the most significant factor in species survival?" were answered with plausible sounding but assumptive responses, ignoring the inherent vagueness.
- Claude: The system produced elaborated and more detailed answers than Gemini. In addition, sometimes it tried to handle ambiguity by offering different ways of understanding. Although in detail, the responses from Claude gave an appearance of completeness, many times the system did not identify ambiguous phrasing and flag it explicitly. For example, given a chemistry question that was somewhat ambiguous, it assumed a particular context without any acknowledgment of a lack of clarity.

- Llama: Compared to Gemini and Claude, handled ambiguity better in cases when questions included complex terminologies or premises with many facets. For example, in biology, Llama sometimes pointed to the requirement of more information. While it flagged ambiguity in some cases, it tended to answer anyway, leading to wrong conclusions. It was inconsistent, especially in physics, when questions needed logical coherence.
- Perplexity: In most cases, it had an easier time than the other models identifying when mathematics and physics questions were ambiguous. It did this by not directly answering questions like "If X + Y =?", which is the higher factor?", but instead flagged them because the data was incomplete. This contrasted with its inability in most instances to recognize an ambiguous premise from a straightforward one in biology and chemistry; it confidently provided answers without enough context.

2. Discipline-Specific Performance

- Biology: All models did relatively better for this discipline since questions in biology often are based on fact recall rather than complicated reasoning. However, even open-ended prompts like "Which organism is the dominator of an ecosystem?" were confidently answered by all models without flagging that more context is required.
- Chemistry: Overall performance was moderate on all the models. Whereas Llama and Gemini provided more consistent responses, in many questions, Claude's elaborate answers concealed its failure to address ambiguous or faulty premises. Perplexity performed reasonably well but slipped into overconfidence from time to time.
- Physics: This was the most difficult subject for all models. Llama and Perplexity did relatively better, occasionally detecting contradictions or ambiguities in questions. For instance, Perplexity detected a faulty premise in "If a car accelerates to -5 m/s, what is the time taken?" while Gemini and Claude attempted calculations.
- Mathematics: Perplexity worked best for ambiguous mathematical instructions by usually complaining about the lack of data. For example, it complained of incomplete inputs such as "Solve for X when 2X + Y = 10." Llama was the next best, while Gemini and Claude mostly assumed arbitrary values or conditions to give an answer.

3. Insights on Ambiguity Handling

- Consistency: Perplexity and Llama were a bit more consistent in handling ambiguity, though their overall success rates were not radically higher than Gemini or Claude.
- Context Awareness: The application of none of the models evidenced robust context awareness. Sometimes, even Perplexity, which may flag ambiguities, often defaults to creating responses without probing for clarity.
- Discipline Impact: Ambiguity handling was most problematic in disciplines like physics and mathematics, where logical reasoning is crucial. Biology and chemistry posed fewer challenges due to their reliance on factual recall.

3. “Are there common patterns in LLM errors for certain question types?”

Observing the dataset I built, the various types of errors the LLMs made were indicative of several repeating patterns:

- Overconfidence in Faulty Responses: Too many times, models provided responses with much confidence for either nonsensical or impossible questions. Such overconfidence can deceive users into accepting those wrong answers as valid.
- Failure to Flag Ambiguity: LLMs almost never acknowledged the information in a question that was missing or not clear. They always attempted to give an answer based on assumptions, showing poor ambiguity detection.
- Inadequate Logical Reasoning: Questions that called for embedded levels of reasoning, like combining ratios or determining a contradiction, were often incorrectly answered. This suggests that LLMs have difficulty in ensuring logical coherence within a response.
- Context Blindness: Each question was processed by LLMs in isolation from relevant science or real-world knowledge, which might have assisted it in identifying the fault.

These patterns show that LLMs are optimized to produce text that sounds plausible rather than critically considering the input for validity and coherence.

4. "Can Interaction, such as Follow-Up Clarification Prompts, Improve the Ability of LLMs to Handle Faulty Science Questions?"

In an experiment done with the Gemini model, interaction through follow-up clarification has yielded quite improved performance by LLMs for faulty science questions. With the follow-up prompt to reconsider its previous answer, Gemini provided a scientifically valid answer by correcting the flawed response, further solidifying that interactive engagement will help refine the performances of LLMs.

The following question was then asked from Gemini: "What is the boiling point of water at sea level when atmospheric pressure is zero? " The first response of the model was wrong; it had confidently told that under zero atmospheric pressure, water would boil at slightly below 100°C or 212°F. This answer showed a lack of understanding of the concept that boiling is essentially dependent on atmospheric pressure.

This response was a perfect example of the overconfidence of LLMs when the premise is flawed. Specifically, the follow-up clarifying question was, "If the atmospheric pressure is zero, then isn't that opposite to the very definition of boiling? " On being asked this, Gemini had to revise its explanation. In the revised response, the model correctly explained that boiling essentially requires atmospheric pressure for the vapor pressure of a liquid to equilibrate with that of its surroundings.

Under zero atmospheric pressure, water would not boil in the usual manner but would go directly into a gas phase through either sublimation or evaporation. This shift in the model's output is indicative of interaction leading to logical re-evaluation and enhanced reasoning in LLMs.

Impact of this interaction –

- Improved Logical Reasoning: The clarification request made Gemini revisit its assumption about boiling and atmospheric pressure. It corrected its mistake by making its explanation consistent with scientific principles. This underlines the fact that interaction can make up for an LLM's initial inability to detect inconsistencies in flawed questions.
- Better Explanation Quality: In the feedback, Gemini gave a more elaborative explanation of the scientific phenomenon, even including terms such as "sublimation" and "vapor

pressure." This shows that interactive engagement may result in richer, more nuanced outputs and hence provides better tools for explaining complex ideas.

- Overconfidence Reduction: The first response from Gemini was overly confident for an answer that was based on a flawed premise. Following the clarification prompt, the model took a step back and was more cautious and correct to admit the logical contradiction in the question. This reduces the possibility of LLMs misinforming users with their confident-incorrect answers.

My recommendation is that to fully realize the power of interactive engagement, LLMs should have inbuilt mechanisms to seek clarification proactively when fed with ambiguous or logically flawed input. Training on datasets designed for multi-turn conversations could also help improve their ability to resolve issues through dialogue. The addition of user-guided verification layers may also help LLMs confirm assumptions before generating a final response.

5. **"Does the complexity of the faulty question, such as multi-step reasoning or subtle semantic faults, make any difference in LLM performance?"**

Yes, the more complex it is, the more LLM performance will be influenced. In this context, complexity is introduced either by the question structure itself, such as multi-step reasoning, or by subtlety of fault, semantic misuse of terms. An analysis of the responses of LLMs to faulty science questions reveals that the more layered and complex the fault, the more the models are bound to fail. This pattern underlines critical gaps in LLM reasoning and interpretative capabilities, especially when moving beyond surface-level understanding.

- Simple Faults: LLMs perform better on simple faults like basic ambiguity. For example:
  Question: "What contributes most to nitrogen fixation: plants or soil bacteria?"
  Response: "Plants contribute significantly," which reflects assumptions but addresses the question.
- Complex Faults: Questions involving multi-step reasoning or combined fault types mostly end in failure. Example:
  Question: "A car accelerates to a speed of -10 m/s. How long did it take?"
  Response: Models either attempt nonsensical calculations or fail to recognize the logical contradiction582-final-doc.

LLMs like Perplexity are slightly better at flagging ambiguity but struggle with multi-step reasoning. Gemini and Claude fail to detect subtle semantic faults, treating incorrect premises as valid582-final-doc. LLMs require superior mechanisms for handling layered logic and complex contradictions; maybe advanced training datasets will bring focus to multi-step reasoning.

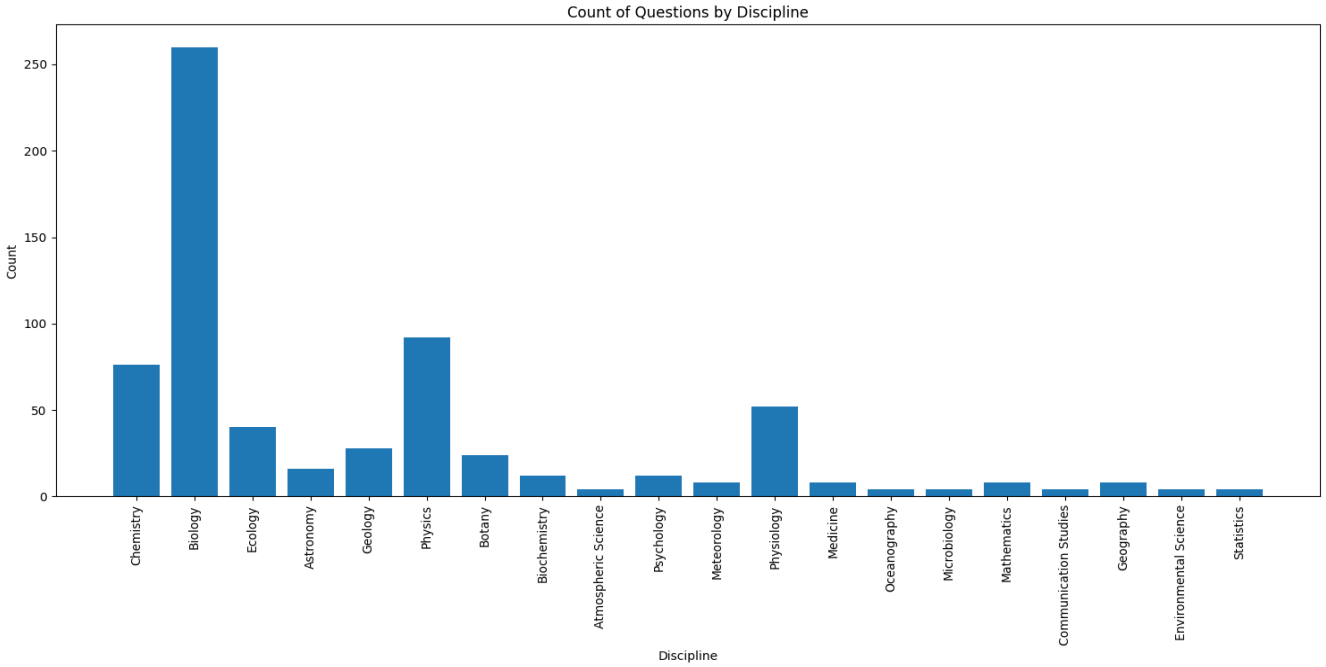
Dataset Description –

The dataset consists of five columns, each capturing important information about the questions and LLM responses:

- Which discipline does the question belong to: The area of study to which the question pertains, such as biology, chemistry, physics, and mathematics.
- Question: The flawed science question developed for testing the LLM.
- Reason you think it is faulty: A short description of why the question is flawed.

- LLM Tested: The LLM tested, such as Gemini or Claude.
- LLM Response: The complete response as given by the LLM under test.

The dataset contains 166 questions, categorized across many scientific disciplines:



The above graph provides a visual representation of the distribution of questions in the dataset across the various scientific disciplines. The said distribution reflects a great effort to make sure the dataset is diverse and wide-ranging in fields and questions' levels of complexity and reasoning required. The dataset ranges over 18 diverse disciplines, including core sciences such as Biology, Chemistry, and Physics; applied sciences like Environmental Science and Medicine; and interdisciplinary areas such as Meteorology and Communication Studies. Such breadth ensures that the dataset captures a wide array of question types and fault scenarios, providing a robust evaluation of LLMs across domains.

The diverse disciplinary representation ensures that a range of fault types is included: for example, logical inconsistencies in Physics, terminological misuse in Chemistry, and ambiguous phrasing in Biology and Medicine. These fault types correspond to the main challenges faced by LLMs in real-world situations. While the number of questions is balanced across disciplines, it stresses fields in which LLMs are frequently challenged interpretively; thus, the set maintains a mix of questions that test both factual recall and reasoning. This ensures that the questions are both scientifically relevant and challenging. Stronger disciplines-for instance, Biology and Physics-point to areas where reasoning or nuances of understanding are usually crucial. Disciplines with fewer questions, such as Statistics or Environmental Science, introduce domain-specific challenges and keep the dataset holistic and relevant.

This dataset is derived from the SciQ dataset hosted on Hugging Face; a repository of science questions commonly used to benchmark AI systems. About 300 questions were selected, and the following steps were performed in creating the final dataset:

1. **Selection of Questions:** Questions were selected from various scientific disciplines to ensure diversity in content and difficulty.

2. **Fault Introduction:** The faults were introduced by prompting GPT-3 with logical inconsistencies, ambiguous language, and incorrect premises. Example:
    - Original: "What is the boiling point of water at sea level?"
    - Faulty: "What is the boiling point of water at sea level when the atmospheric pressure is zero?"
3. **Faults Validation:** Each modified question was reviewed by me to ensure that the introduced faults were meaningful and would challenge the reasoning abilities of LLMs.
4. **Testing on LLMs:** Questions were individually submitted to leading LLMs, such as Gemini and Claude, and their responses were recorded. The testing environment ensured consistency across models by me.
5. **Annotation:** The reason each question was considered faulty was documented for each question in a separate excel sheet, to contextualize the analysis.

Fault Types:

Faulty science questions were carefully designed to exploit reasoning and interpretative weaknesses in state-of-the-art Large Language Models (LLMs). Each question falls into one of three primary fault categories, as described below:

1. **Ambiguity:** Questions in this category intentionally use unclear, ambiguous, or incomplete wording to test an LLM's ability to recognize ambiguity and seek clarification. Examples include the following:
    - *"Which contributes more toward nitrogen fixation-plants or soil bacteria?"*
      Fault: The word "more" is not defined with regard to measures or contexts.
    - *"Which organism is considered dominant within an ecosystem?"*
      Fault: The question is too open without defining "dominance" population, biomass, etc.

   LLMs frequently do not flag ambiguity and respond with plausible but assumptive responses. Ambiguous prompts tended to result in overconfident answers rather than requests for clarification, especially in Biology and Chemistry. This type of fault shows that it is necessary for an LLM to recognize unclear inputs and interactively ask follow-up questions.
2. **Logical Contradictions:** These questions have premises or conditions that are against scientific principles or real-world logics, challenging LLMs' reasoning. Examples include the following:
    - *"What happens to water when it is heated to 0 Kelvin?"*
      Fault: The premise is physically impossible since 0 Kelvin is absolute zero.
    - *"A car accelerates to a speed of -10 m/s. How long did it take?"*
      Fault: Negative speed is meaningless in this context.

   Logically inconsistent inputs were particularly hard for LLMs in Physics and Mathematics. However, Perplexity and similar models showed better detection of these inconsistencies but still tried to produce many nonsensical answers. This class of fault underlines the capability issues of LLMs to enforce logical coherence and spot serious inconsistencies in inputs.

3. **Misuse of Terminology:** Questions make deliberate misuse of scientific terms or concepts to test LLMs on whether they can catch those errors in terminology. Examples:
    - *"Sublimation happens when water reaches its boiling point."*
      Fault: The process of sublimation involves the direct phase change from solid to gas, which is not boiling.
    - *"What is the chemical reaction called evaporation during photosynthesis?"*
      Fault: Photosynthesis involves no evaporation; it was the deliberate misuse of terms.

   Subtle terminological errors usually remained undetected, with LLMs responding confidently as if the terminology was correct. The questions on chemistry were the most prone to this type of fault. The inability to detect misuse of terminology underlines certain gaps in the semantic understanding of LLMs and reliance on pattern recognition rather than context-sensitive reasoning.

A small snippet of my data set:

Discipline	**Question**	**Faulty Reason**	**LLM**	**Response**
Biology	"What contributes most to nitrogen fixation: plants or soil bacteria?"	Ambiguity in comparative context	Gemini	"Plants contribute significantly."
Chemistry	"What happens to water when it is heated to 0 Kelvin?"	Logical impossibility (0 Kelvin is absolute zero)	Claude	"Water freezes completely."
Physics	"A car accelerates to a speed of -10 m/s. How long did it take?"	Negative speed is physically meaningless	Gemini	"Acceleration depends on time given."

<u>Experimental Design</u> –

This experiment compared the performance of four state-of-the-art Large Language Models on 166 faulty science questions. Each LLM was a state-of-the-art model concerning language understanding and reasoning competence. Their performance was thus analyzed based on their ability to identify faults in questions and providing appropriate responses. The models included Gemini, Claude, Llama, and Perplexity. Every question in this dataset was intentionally crafted to either have a logical inconsistency, ambiguous phrasing, or an incorrect premise. The main aim was to see whether LLMs would be able to identify the faults in these questions. Testing was structured as follows:

1. **Input of Questions:**
    - All 166 questions were given as input, one by one, to all four LLMs under the same experimental settings.

2. **Recording of Responses:**
    - Each response given by an LLM was recorded verbatim in the dataset.
    - Each response was assessed for:
        - Whether the LLM correctly identified the fault in the question.
        - Whether the answer provided by the LLM was correct or not.

- The clarity of the response in ambiguous cases.

3. **Detection of Fault:**
   - If the LLM indicated the fault in the question explicitly, for example, it finds logical inconsistency or ambiguity; such cases were marked as successes.

4. **Performance Evaluation:**
   - Responses were categorized as:
     - Correct-Fault Identified: The LLM correctly flagged the fault and did not give an incorrect answer.
     - Incorrect Answer: The LLM gave a response but did not identify the fault.

Steps Taken to Test and Compare Responses

1. A total of 166 questions were carefully developed to ensure that fault types were evenly distributed across the disciplines of biology, chemistry, physics, and mathematics.
2. All questions were in the same format to reduce any form of bias during LLM processing.
3. Each question was subjected to all four LLMs, whose responses were captured in the dataset for analysis.
4. Responses were coded as follows: Incorrect answer, Correct Answer.
5. From the recorded data, some performance metrics were computed quantifying the ability of each model in detecting faults and dealing with ambiguities.

Results and Analysis –

The analysis of the dataset and LLM responses revealed several key insights into the performance and limitations of current large language models when confronted with faulty science questions:

1. Variability in Performance Across Disciplines
   - Strengths: LLMs like Gemini and Claude demonstrated higher success rates in biology and chemistry questions, particularly for those that required factual recall rather than complex reasoning. These models often performed well when the questions contained explicit context clues or were fact-based despite the introduced faults.
   - Weaknesses: Physics and mathematics emerged as the most challenging disciplines. Questions involving logical inconsistencies, ambiguous premises, or unit mismatches frequently led to incorrect or overconfident responses. This suggests that LLMs struggle more with reasoning-based queries compared to factual recall.

2. Common Patterns in LLM Failures
   - Overconfidence in Answers: One of the striking trends in all LLMs was that they tended to give confident but wrong answers, even when faced with glaringly faulty premises. For instance, questions with logical impossibilities were often answered as if they were valid, without any acknowledgment of the fault.

- Ambiguity Misinterpretation: Several LLMs fell prey to not recognizing when phrases could be ambiguous or where further context might be required; rather than seeking an explanation, they would attempt plausible-sounding answers that were very much further from the truth.
- Logical Errors: Questions with multi-step reasoning or those that require an understanding of conflicting premises often revealed weaknesses in the LLMs. For example, when questions combined contradictory ratios or units, models did not recognize the flaw and completed the calculation based on incorrect input.

3. Differences in Model Behavior

Model	**Strengths**	**Weaknesses**	**Best-performing Disciplines**	**Worst-performing Disciplines**
Gemini	Strong factual recall in Biology and Chemistry.	Struggled with logical reasoning and ambiguity detection.	Biology, Chemistry	Physics, Mathematics
Claude	Elaborate and detailed responses.	Failed to explicitly flag ambiguity and often masked faults with verbose answers.	Biology, Chemistry	Physics, Mathematics
Llama	Better at identifying ambiguity and logical faults than others.	Inconsistent performance in handling multi-step logical reasoning.	Physics, Mathematics	Biology, Chemistry
Preplexity	Strongest in identifying ambiguities and incomplete data.	Limited ability to detect logical contradictions or subtle semantic faults.	Physics, Mathematics	Biology, Chemistry

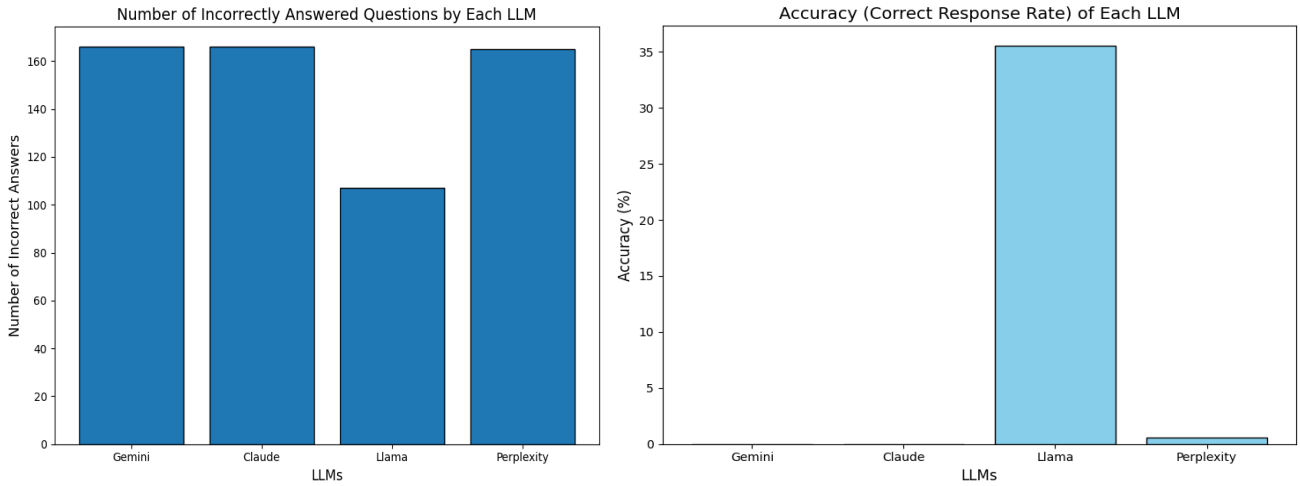
4. Types of Faults Which Were Most Difficult
    - Logical Contradictions: Questions framed by contradictory premises, such as "A vehicle accelerates to a negative speed of 10 m/s", often evade the fault detection mechanisms in LLMs.
    - Ambiguous Prompts: This mostly came from open or vague questions with too little context for specific responses. Models were very rarely flagging these as ambiguous but attempting to give some specific answer.
    - Subtle Semantic Faults: The questions that had subtle linguistic traps included misuses of scientific terms, and LLMs were very poor at handling them. For instance, using the term "evaporation" instead of "sublimation" in a chemistry context often led to an incorrect response without the model recognizing the misuse.

Accuracy and Number of Incorrectly Answered Questions

Among them, Llama performed much better, with a correct response rate of 35.54%, meaning it had the strongest reasoning or accuracy capability among the tested LLMs. In cases where Gemini, Claude, and Perplexity failed to answer questions correctly, significant weaknesses in reasoning or ambiguity handling were indicated. This is further manifested in their inability to spot or resolve faulty or ambiguous questions effectively. These models may heavily rely on pattern matching or factual recall rather than deeper reasoning capabilities. They may not have

mechanisms for flagging faulty logic or ambiguity and instead attempt to answer questions as if they were valid.

On most occasions, Llama consistently outperformed others on accuracy and on giving more incorrect answers, and that makes Llama a very reliable model for this dataset. Probably, the model has gone through a varied dataset, or probably a tougher dataset while training. Probably its inner mechanism is much better to tackle faulty premises or even questionable questions. While Perplexity improved marginally over Gemini and Claude, it did not differ significantly in its overall performance. High error rates across all models except Llama indicate a general problem of LLMs with problematic science questions. This calls for significant improvement in reasoning, ambiguity handling, and fault detection.



Some key limitations brought out by the dataset are enumerated below:

1. Inability to Detect Faulty Premises:
   - Logical Inconsistencies: LLMs failed to flag impossible scenarios, such as negative speed, and embraced implausible premises, including misuses of scientific terms.
   - Implication: Absence of mechanisms for logical soundness verification or aligning responses with scientific knowledge.
2. Difficulty Handling Ambiguity:
   - Insufficient Clarification: The models answered questions that were vague or incomplete without attempting to seek more information.
   - Misinterpretation: Ambiguous language frequently led to incorrect responses.
   - Implication: Poor handling of nuanced context and incomplete inputs.
3. Overconfidence in Incorrect Answers:
   - Confident but Wrong: LLMs frequently gave definitive answers to nonsensical questions.
   - No Uncertainty: Models rarely conveyed doubt, misleading users.
   - Implication: Overconfidence can propagate misinformation.
4. Insufficient Logical Reasoning:
   - Multi-Step Reasoning: Models struggled with questions requiring layered logic or contradiction detection.

- Implication: Inability to maintain logical coherence.

5. Limited Contextual Understanding:
   - Ignoring Real-World Knowledge: Responses sometimes contradicted basic scientific principles.
   - Context Isolation: Models failed to relate questions to broader scientific concepts.
   - Implication: Lack of grounding in real-world knowledge.
6. Variability Across Models:
   - Inconsistent Performance: Different models showed varying weaknesses (e.g., ambiguity vs. logical faults).
   - Implication: No model demonstrated comprehensive reasoning strength.
7. Inadequate Error Detection:
   - No Self-Awareness: Models did not recognize their lack of information or accuracy.
   - No Clarification Prompts: Failure to ask follow-up questions limited their interactive utility.
   - Implication: Reduced reliability in identifying faulty inputs.
8. Failure to Provide Safe Alternatives:
   - Misinformation Risk: Confidently wrong answers raised ethical issues.
   - Missed Educational Opportunities: Models failed to teach users about faults or misconceptions.
   - Implication: Lost opportunities for safe and pedagogically useful interactions.

Conclusion –

This study underlines substantial issues when Large Language Models face faulty science questions designed to expose them to reasoning weaknesses. Among the key findings is the systematic failure across all fault types, especially in logical contradictions, dealing with ambiguity, and terminological misuse. Physics and Mathematics were the toughest domains and thus revealed weaknesses of the LLMs in the realms of multi-step reasoning and coherence, while Biology and Chemistry seemed to be closer to a relative strength in simple factual recall.

The dataset, which ranges over 18 disciplines with very diverse faults, serves as a strong benchmark to appraise the reasoning of LLMs and their fault detection capability. By systematically analyzing model behavior across fault types and disciplines, this study highlights critical areas for improvement that need attention, including enhancement of context awareness, development of mechanisms for ambiguity detection, and reduction in overconfidence in faulty answers.

While current LLMs demonstrate remarkable language generation capabilities, their inability to handle flawed premises is one of their major weaknesses. Understanding these weaknesses is important in making these models safe and effective to apply in real-world situations where the highest degree of precision and logical stringency is called for.