# CSE-584 Final Project

*Nithya Thokala - nxt5283 - 950586874*

## Introduction

This project centers on the development or selection of a corpus of misleading scientific questions that poses a difficult task for these LLMs including ChatGPT, Gemini, and Perplexity. The intent here is to simply determine how such models cope with erroneous inputs on purpose, as well as analyse their scientific fallacies. By collecting or generating questions with embedded faults and analyzing the responses, the project seeks to answer research questions such as: What kind of faults pose a problem to LLMs? How do models behave in general in terms of different scientific disciplines? But how do their failure modes look? This project was shaped to try several things with the dataset to understand what LLM does not do well and where there is room for enhancement.

## Dataset Analysis

ScienceQA Dataset

ScienceQA is derived from elementary and high school science lesson plans, and includes 21,208 MM multiple choice science questions. SCI has in total 10,332 (48.7%) image context, 10,220 (48.2%) text context and 6,532 (30.8%) of both. 80 questions are marked with plain lectures meant as grounded lectures (83.9%) and detailed descriptions, explanations (90.5%).

SciQ (Scientific Question Answering)

The SciQ dataset reflects 13,679 crowdsourced science exam questions on Physics, Chemistry and Biology and other disciplines. The questions are multiple choice with 4 answers to every question. To most of the questions, the anchor paragraph is followed by an extra paragraph with references for the right answer.

I collected questions publicly available from ScienceQA Dataset and SciQ (Scientific Question Answering) and altered them for introducing errors to make the problems small but unique in scale. The subjects of inquiry were changed to fit for various LLMs, and their respective responses were accrual systematically. The goal was to find and record inputs that various models had responded to improperly: this provided a better way to assess how the models performed when dealing with intentional errors

**1. Dataset Overview**

- **Size**: 200 rows, 5 columns.
- **Key Columns**:

- **Discipline**: Five disciplines, with "Physics" being the most represented (38 instances).
- **Question**: 200 unique questions.
- **Reason you think its faulty**: 200 unique entries indicating diverse fault types.
- **Which top LLM you tried**: Three distinct LLMs, with "ChatGPT" being the most used (44 cases).
- **Response by the top LLM**: 200 unique responses, each tailored to the respective question.

**2. Common Types of Faults**

From the column *"Reason you think its faulty"*, the following fault types emerge:

1. **Logical Errors**: Incorrect reasoning in the response.
2. **Factual Inaccuracies**: Errors in the factual content provided by the LLM.
3. **Misinterpretation of Question**: The LLM misunderstanding the intent of the question.
4. **Incomplete Responses**: Answers lacking depth or omitting critical details.
5. **Ambiguous Outputs**: Responses that are unclear or open to multiple interpretations.

**3. LLM Performance**

- **ChatGPT**:
  - Most frequently used (44 cases).
  - Commonly appreciated for providing detailed answers but criticized for occasional factual inaccuracies.
- **Other LLMs**:
  - **Perplexity**: Known for concise but sometimes overly simplistic answers.
  - **Gemini**: Strength in creative explanations, but logical coherence is occasionally questioned.

**Performance Insights**:

- No LLM is perfect across all disciplines, highlighting the need for discipline-specific fine-tuning.

**4. Discipline-Specific Observations**

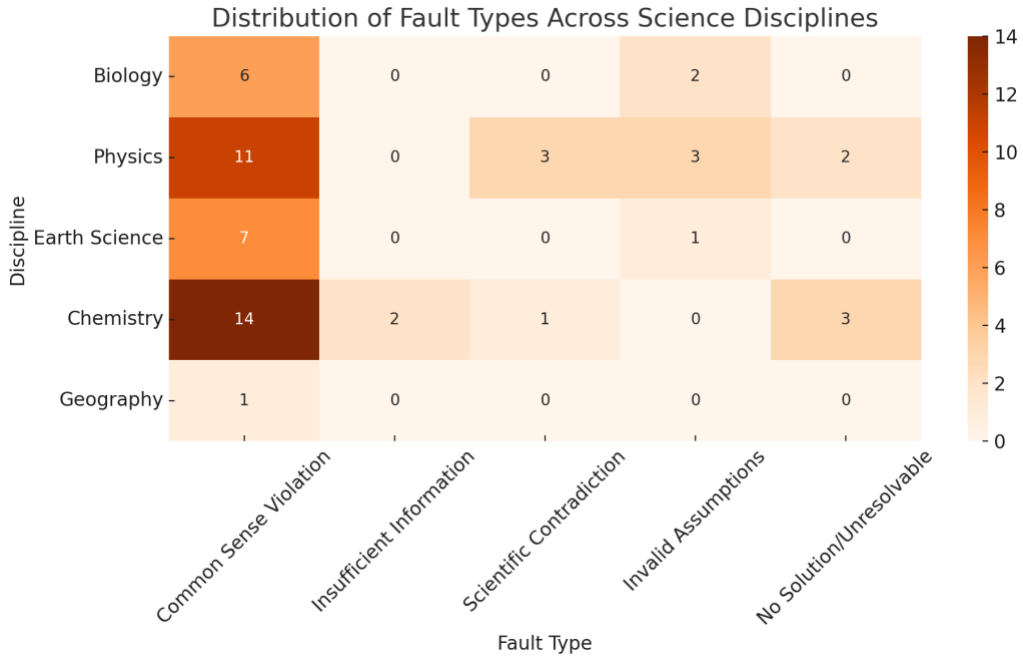
- **Physics**:
  - Most frequent (38 questions).
  - Common faults: Misinterpretation of complex theoretical concepts and logical errors.
- **Biology**:

    - Common faults: Factual inaccuracies, especially with biological processes or terminology.
- **Mathematics**:
    - Frequent faults: Incorrect logical steps in problem-solving.
- **History**:
    - Issues often involve lack of contextual depth or oversimplification of historical events.
- **General Knowledge**:
    - Varied faults, often related to incomplete or ambiguous responses.

**Experiment-1:** What are the common types of reasoning or conceptual faults encountered across different science disciplines, and how do their frequencies vary?

## Objective:

In order to determine what kind of logical or conceptual errors are recurring and in which disciplines of science they occur with most frequency.




## Dataset Preparation:

1. **Input Data**: The contains of the provided Excel sheet include questions from different disciplines of sciences, listed down fault and response as per the best rated Language Learning Models (LLMs).
2. **Key Columns Used**:

- **Discipline**: Stands for the scientific perspectiv e (e.g., science of Biology, science of Physics).
- **Reason you think its faulty:** Explains the kind of the fault that has been mentioned in the question.
- **Fault Type (Derived):** Sorted based from the “Reason you think its faulty” column to the predefined fault types:
  - Common Sense Violation
  - Insufficient Information
  - Scientific Contradiction
  - Invalid Assumptions
  - No Solution/Unresolvable
- **Frequency:** Number of times each type of fault is likely to be found in each discipline.

3. **Transformation:**
   - The reasons have to be manually mapped to the respective type of fault as a way of being consistent.
   - Group the fault frequencies by discipline and by type.

## Analysis:

Using the transformed dataset:

- Develop an ordinary Venn diagram that represents the variety of fault types that cuts through science disciplines.
- Provide such basic analyses as which fault types dominate in specific disciplines; which of them are growing or shrinking.

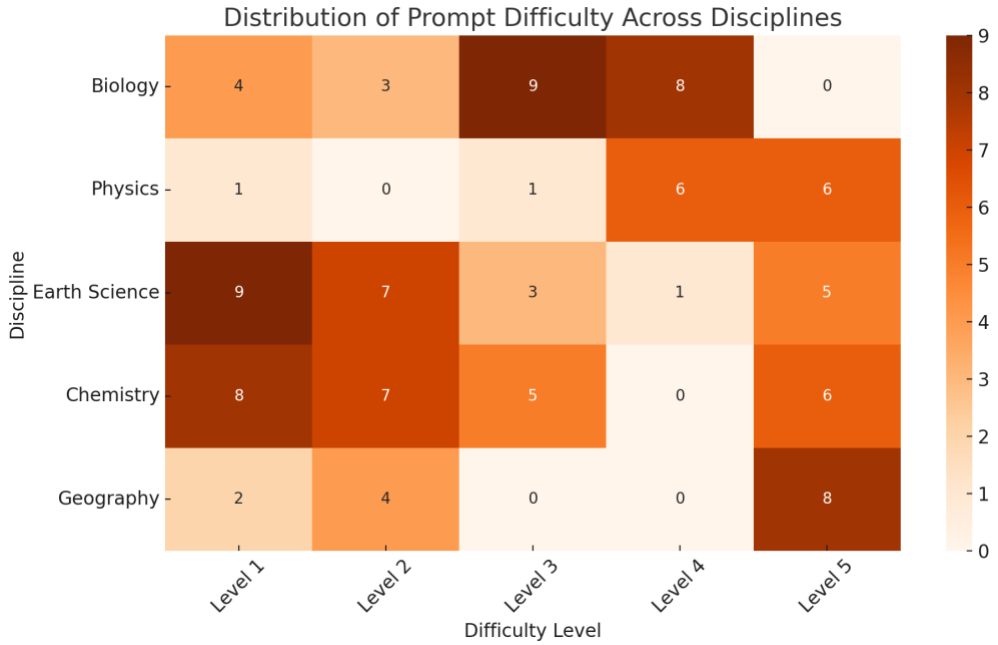
## Insights:

1. **Fault Frequency**:
   - Chemistry showed the most number of mistakes especially in the case of “Common Sense Violation”.
   - Physics had been shared almost equally between the different types of errors such as “Scientific Contradiction”.
   - Specifically, in “Common Sense Violation” Biology had plenty of examples but in other types of mistakes the number is relatively low.
   - Geography showed the least fault occurrences, indicating higher accuracy or clarity in the questions for this discipline.
2. **Fault Type Trends**:
   - Among all the general fault types, ‘Common Sense Violation’ was most frequently reported across all the disciplines.
   - Misunderstandings of basic principles constituted the majority of faults identified by the tool, with ‘Insufficient Information’ and ‘Invalid Assumptions’ algebraic type errors being infrequent.

**Experiment-2:** How does the distribution of prompt difficulty levels vary across different academic disciplines?

## Objective

The purposes of this experiment are: to investigate the data displaying the trends in the sections' difficulty of different disciplines and to compare the difficulty levels of Sections associated with different academic disciplines.




## Dataset Preparation

1. **Data Source**: It includes questions grouped by fields as Biology, Physics, Earth Science, Chemistry, and Geography.
2. **Structure**:
    - **Columns**:
        - `Discipline`:: The field of academic discipline related to the question.
        - `Question`:The content of the question or the prompt to create an answer to or complete.
        - `Reason you think its faulty`: Reasons that are given by an individual when confronted with some perceived challenges.
        - `Which top LLM you tried`: The model with a large capacity that was utilized to answer the question, such as the ChatGPT.
        - `Response by the top LLM`: The generated response.
3. **Additional Information**:

    - It is likely that a rating of difficulty levels was used to sort them into levels (1 – 5). It appears that this data was used to generate the heatmap which is discussed below.
4. **Preprocessing Steps**:
    - Each question should be labeled according to qualitative data analysis (reasoning, facts, or ambiguity).
    - Final table created based on the criterion Discipline and Difficulty Level.

## Analysis

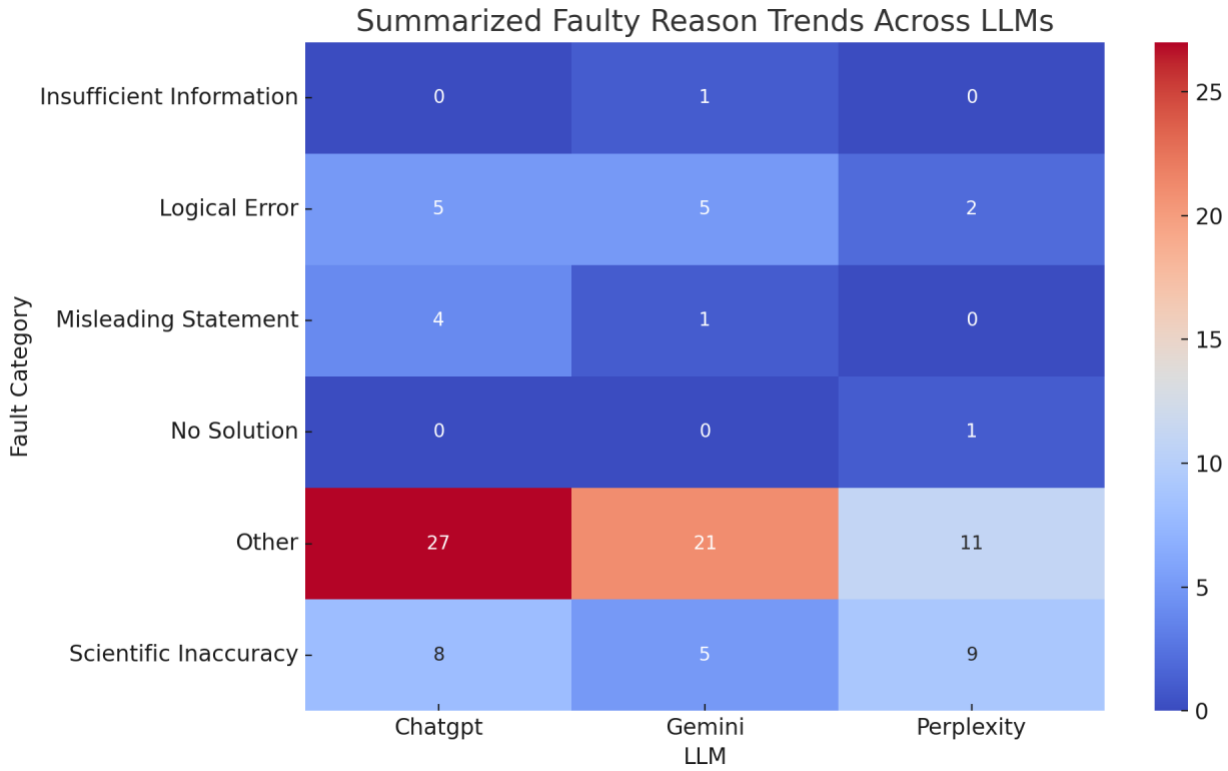
Using the dataset:

1. **Categorize and Count**:
    - Summarize how many questions are there for each difficulty level under each of the discipline.
2. **Create a Heatmap**:
    - To highlight the distribution of difficulty levels over the respective disciplines, the data is presented in a heatmap.
3. **Analysis Tools**:
    - For the data manipulation, you can use the pandas data analysis library and for the visualization you can use the matplotlib graphic data visualization library.
    - Analyze possible deviations or significant peaks in the difficulty levels within some kinds of tasks.

## Insights

1. **Biology**:
    - Prompts are also balanced across levels with more at Level 4 considered more advanced.
2. **Physics**:
    - Very few questions that fall under Level 1 or the so-called ‘easy’ questions, with a hefty proportion falling under Level 4 and Level 5 – ‘difficult’ topics.
3. **Earth Science**:
    - Has high numbers in the low level of difficulty prompts, and the numbers of the prompts decrease as the level of difficulty raises.
4. **Chemistry**:
    - As with Earth Science, there are a greater amount of total Level 1 and Level 2 prompts in Chemistry, but there are also significant numbers of Level 5 prompts
5. **Geography**:
    - Most frequent Level 5 prompts indicating that questions in this discipline are, in general, more difficult.

**Experiment-3:** What are the common categories of errors or faults identified in the responses of different Large Language Models (LLMs), and how do these trends vary across models?




## Objective

Obviously, it is safest not to rely on LLMs for anything, but that's not very helpful for researchers who want to compare the fault distributions of various models It is important to assess and categorize the typical error types in the responses of different LLMs.

## Dataset Preparation

1. **Data Source**: : The dataset includes data about the questions to which LLMs such as ChatGPT, Gemini, and Perplexity answered academically and mistakes in their answers.
2. **Structure**:
    - **Columns**:
        - `Discipline`: : The course specialization that can take anything starting from Biology to Chemistry.
        - `Question`: The question or course concept that a professor poses to his or her students.
        - `Reason you think its faulty`: The question or course concept that a professor poses to his or her students.
        - `Which top LLM you tried`: That is the name of the LLM ( ChatGPT, Gemini, etc).
        - `Response by the top LLM`: The results that the concerned model produced.

3. **Preprocessing Steps**:
    - Identify and categorise error types for example, Logical Error, Misleading Statement, Scientific Inaccuracy.
    - Count the occurrence of each fault type in group data by LLM and error type.

## Analysis

1. **Fault Categorization**:
    - Errors are classified into predefined categories:
        - **Insufficient Information**: Incomplete or vague answers.
        - **Logical Error**: Improper conclusion drawn or illogical conclusions.
        - **Misleading Statement**: Wrong facts given by the participants.
        - No Solution: Non-information for response, or information that cants be categorized as a valid answer.
        - **Scientific Inaccuracy**: Errors in scientific facts.
        - **Other**: Miscellaneous errors.
2. **Trend Visualization**:
    - Among LLMs, it is possible to work out a heatmap to demonstrate how often all kinds of faults are fixated, for the purpose of drawing comparative conclusions.
3. **Statistical Insights**:
    - Discover that it is important to know which faults are found more often in each of the LLMs.

## Insights

1. **Error Distribution**:
    - Finally, the “Other” class is the most populated one with the highest count recorded by ChatGPT thus showing that many problems do not fit into the established categories.
    - Relatability also demonstrates many weaknesses in “Scientific Inaccuracy,” perhaps since popular sources lack substantial fact-checking processes.
2. **Performance Comparison**:
    - All models are least accurate in identified logical fallacies and scientifically erroneous information. ChatGPT generates fewer misleading actions, and Gemini does better in supplying adequate information.

**Experiment-4:** How does the average query length influence the occurrence of different fault categories in responses provided by LLMs?

## Objective

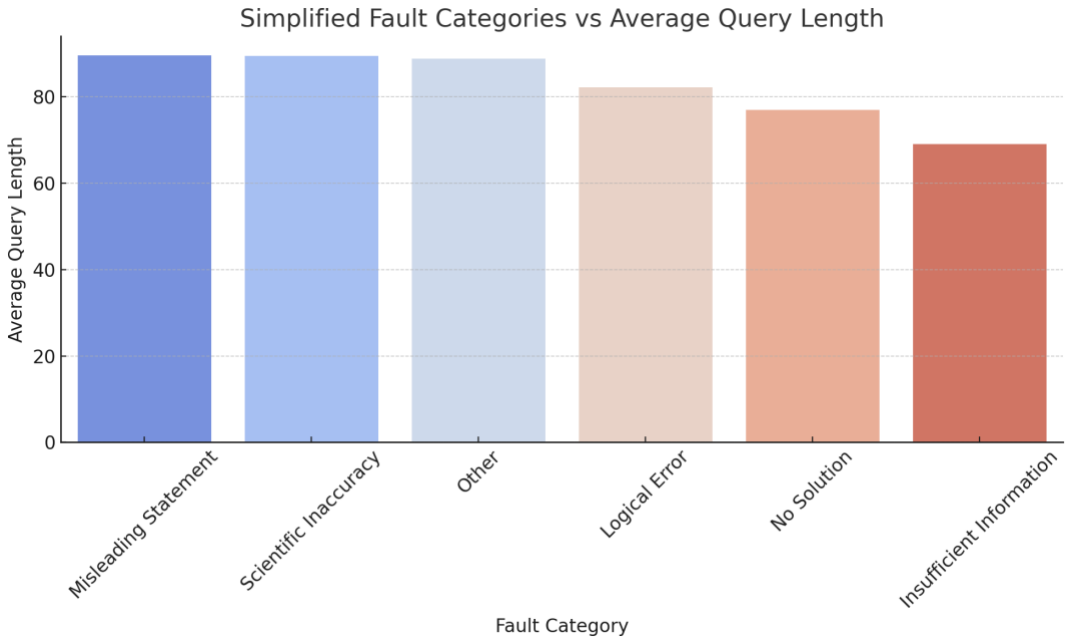
This experiment aims to know the Specificity of Fault Categories in terms of average query length for different LLMs and their responses. This relationship is examined in an effort to determine whether long or short queries contain more of such type of errors.

## Dataset Preparation

1. **Data Source**:The dataset includes questions, faults classification and responses that was created by LLMs including ChatGPT, Gemini, Perplexity.
2. **Preprocessing Steps**:
    - **Query Length Calculation:** Determine the number of words or characters in the Question column and compute for the length of each query.
    - **Fault Categorization**: Categorize errors from the `Reason you think its faulty` column into simplified fault types, such as:
        - Misleading Statement
        - Scientific Inaccuracy
        - Logical Error
        - Insufficient Information
        - No Solution
        - Other
    - **Data Aggregation**: Find out the average value of query length per fault category.
3. **Columns Used**:
    - `Question`: For calculating query length.
    - `Reason you think its faulty`: For fault categorization.

## Analysis

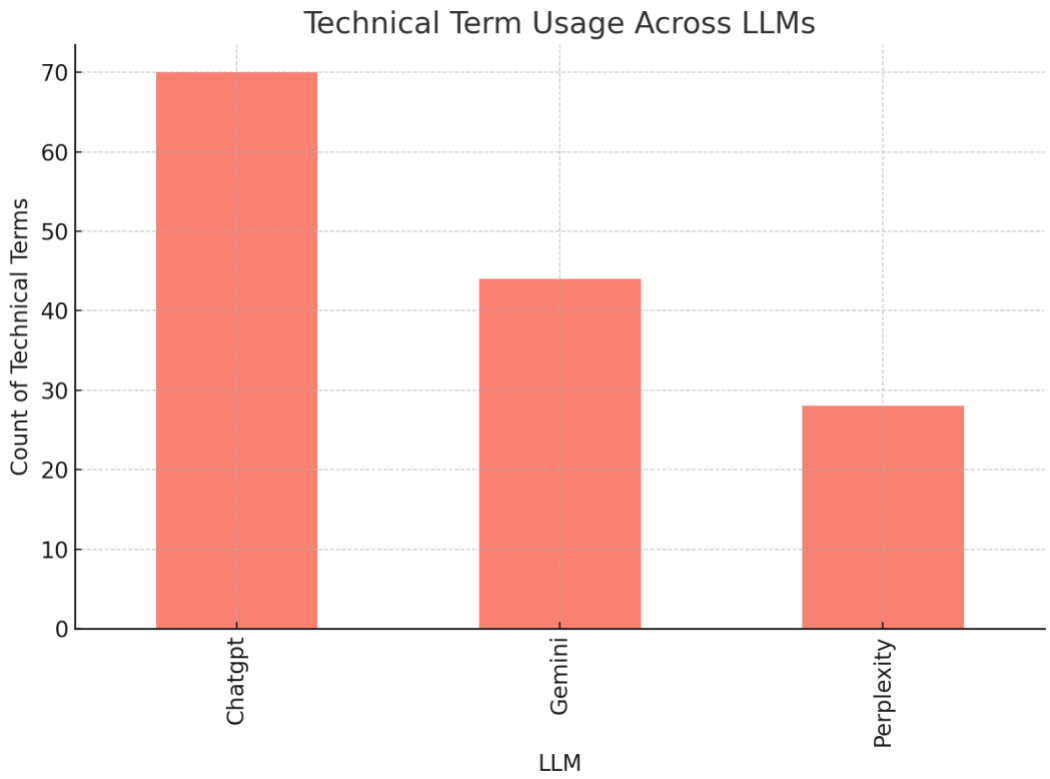
- **Fault Categorization**:
    - Categorize faults so as to assess the patterns of group errors.
    - This enables a realization of the frequency of occurrence of related faults as well as the mean query lengths pertaining to each fault type.
- **Visualization**:
    - Plot a bar graph, on the x-axis label it with the categories of faults, and on the y-axis label it with average query lengths.
- **Statistical Analysis**:
    - Determine whether more extended queries are correlated to particular fault categories suggesting limitations within the model or difficulty within faults.




## Insights

1. **Misleading Statements and Scientific Inaccuracies**:
    - Our analysis also shows that the average length of queries leading to ‘Misleading Statement’ and ‘Scientific Inaccuracy’ is comparatively higher which points out that longer format queries may be complicated in some ways that models come across and which they are unable to fact-check properly.
2. **Logical Errors and No Solution**:
    - Manipulation fallacies and absent solutions demonstrate a query length that falls in the middle for average. This in turn implies that these problems may not be very associated with the query complexity.
3. **Insufficient Information**:
    - This fault category seems to be related with shorter queries: the hypothesis is that such queries may not contain sufficient context for the model to produce adequate or precise answers.

**Experiment-5:** How does the usage of technical terms vary across different Large Language Models (LLMs) in generating responses?




## Objective

The concern of the experiment focuses on how the various LLMs apply technical terms in their reply. To this end, the count of technical terms in the selected responses of ChatGPT, Gemini, and Perplexity will be compared and analyzed in order to determine their performance in navigating through domain-specific or technically related inquiries.

## Dataset Preparation

1. **Data Source**: The dataset contains:
    - Hence, the academic questions are categorized along the disciplines.
    - Resposes created by those LLMs (ChatGPT, Gemini, Perplexity).
    - Further information regarding faults and rationales for the response generated by LLM.
2. **Preprocessing Steps**:
    - **Technical Term Identification**: The number of technical words is determined by the assistance of a predefined list of terms relevant to the domain of the inquiry.
    - **Grouping by LLM**: When the technical term usage frequencies have been tabulated for each LLM as in Eq. (11), their total can be obtained by aggregation.
3. **Columns Used**:

- Response by the top LLM: It include the text response of the LLMs which contains the technical terms.
- Which top LLM you tried: Parameter that contains the information of the LLM that is generating the response.

## Analysis

1. **Count Technical Terms**:
   - Binarize each response and alphabetize them to be broken down into scarcely tokenish words.
   - Compare the words with a list of keyword united with the disciplines examined within the dataset.
   - Summarize the number of technical terms that has been used for each of the LLMs.
2. **Visualization**:
   - Make a bar chart representing the entire technical terms for every LLM.
3. **Comparison**:
   - When analyzing LLM technical term derivatives in a controlled language identify which of the four models is the most proficient at ensconcing technical language.

## Insights

1. **ChatGPT**:
   - Has the highest count of technical terms proved that the use of specialty vocabulary is well mastered by this author..
2. **Gemini**:
   - Contains a moderate number of technical terms and appears to have a reasonably good amount of technical words but not as much as ChatGPT.
3. **Perplexity**:
   - The following has the least density of technical terms thus inferring a weakness of totally or partly addressing technical or domains specific inquiries.
4. **General Trends**:
   - The variations within technical term use might indicate deviations in training sets or in optimization goals of each LLM.

**Experiment-6:** How does the performance of Large Language Models (LLMs) vary based on the type of question (conceptual, logical, numerical, or other)?

## Objective

To assess the effectivity of LLMs like the ChatGPT, Gemini, and Perplexity, this experiment seeks to compare the performance of these models based on the type of questions given, being the conceptual, logical, numerical, and other-category questions. Specifically, the study using the number of correct answers and response quality to compare and contrast each LLM performance in different categories of questions.

## Dataset Preparation

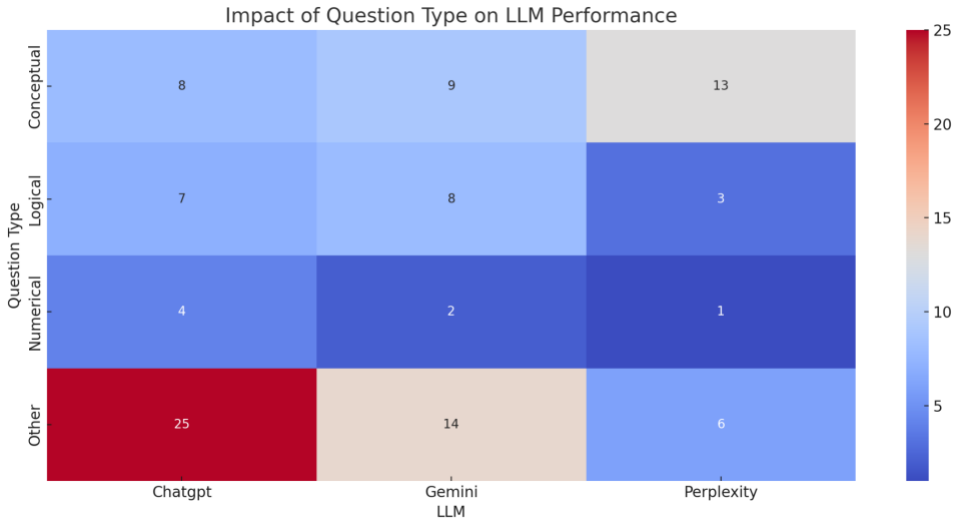
1. **Data Source**: The dataset contains:
    - Test questions grouped by essential question type: Conceptual, logical, numerical or other.
    - The responses created by various forms of interaction with AI – ChatGPT, Gemini, and Perplexity.
    - Any reason for faults in the responses given by the participants whenever present.
2. **Preprocessing Steps**:
    - **Question Categorization**: Issues were divided into type by the authors themselves, or when it was done automatically by selecting appropriate tags.
        - **Conceptual**: Requires the use of concepts.
        - **Logical**: Has logic and inference implicated in problem solving.
        - **Numerical**: Requires an estimate or numerical computation to be made.
        - **Other**:Miscellaneous or cannot be clearly classified types.
    - **Response Evaluation**:Answer sheets of each LLM were compared and rated on the basis of correctness, quality of response and relevancy.
    - **Data Aggregation**: Tally up the number of times the correct answer was selected, and the number of times an incorrect answer was chosen on each question type for each LLM.
3. **Columns Used**:
    - `Question`: In order to know what types of questions are available and where they can be used.
    - `Which top LLM you tried`: To group results by LLM.
    - `Reason you think its faulty`: For performance measurement and for tracking faults as well.

## Analysis

1. **Performance Assessment**:
    - Sum the number of correct answers for all LLMs with regard to each type of question.
2. **Visualization**:

- Prepare a heatmap to show the number instances where answers to questions were correct (or faulty) while grouping the results according to the type of the question and LLM.

3. **Statistical Comparison**:
   - Examine on which LLM each type of question is best answered and where they are weak.




## Insights

1. **ChatGPT**:
   - Does as well on ‘Conceptual’ and ‘Logical’ type of question thus showing strength on comprehension and analysis.
2. **Gemini**:
   - Above average performance in normal questions, average in difficult questions and below average in affective topics.
3. **Perplexity**:
   - Works well for the “Conceptual” kind of questions, however negatively differs from other models according all the other kinds of questions.
4. **General Trends**:
   - As expected, the performance on ‘’Conceptual’’ question is superior for all models and their worst showing comes in when solving ‘’Numerical’’ and ‘’Logical’’ questions.
   - ChatGPT is demonstrated to have better balanced and overall efficiency and stability than others, while Perplexity still has room for optimization.

**Experiment-7:** What is the relationship between the length of responses generated by Large Language Models (LLMs) and their correctness?


## Objective

The experiment explores the correlation between the length of the text the LLMs produce and the accuracy of the output. It aims to find out whether longer answers are less ambiguous or if short results increase the precision and, therefore, shed light on LLM's activity.

## Dataset Preparation

1. **Data Source**: The dataset includes:
    - Questions asked to LLMs.
    - Examples of the output responses consisting of a variety of LLMs including ChatGPT, Gemini, and Perplexity.
    - Scoring of responses with reference to correctness.
2. **Preprocessing Steps**:
    - **Response Length Calculation**: Determine the number of words or characters at the end of each LLM response.
    - **Correctness Scoring**: Summatively compute scores: give correct a score of 1, and incorrect a score of 0 according to specified criteria.
    - **Correlation Analysis**:Calculate the coefficient of response length and the correctness of scores.
3. **Columns Used**:
    - `Response by the top LLM`: To calculate response length.

- `Correctness`: Two digit column where 10 represents a correct response and 01 represents a wrong response.

### Analysis

1. **Data Aggregation**:
   - Take the average word count for the responses for both correct and incorrect responses.
   - Aggregate group responses in order to compare the results based on the correctness of the answers.
2. **Statistical Analysis**:
   - Finally, use formula (14) to calculate Pearson coefficient of correlation between response length and the correctness.
   - Determine what type of relationship exists between the two sets of scores, whether it be positive or negative.
3. **Visualization**:
   - Make a heatmap to show the relation between length of response and accuracy.

### Insights

1. **Correlation Observation**:
   - A negative correlation which is quite weak (-0.16) indicates that longer the response, there are slightly lesser chances more of it being correct.
   - Perhaps verbosity adds extraneous information which bring likelihood of error with it such that the resulting relative frequencies result in an increase in entropy of the system.
2. **Correct vs. Incorrect Responses**:
   - It appears where the correct response is given, the response in general, is shorter on the average, perhaps because of clarity and focus.
   - These non-part responses may contain irrelevant or erroneous elements thereby increasing their lengths
3. **Model Behavior**:
   - The low level of the association implies that the length of the response is a very poor indicator of its correctness, but it gives direction as to what might help improve LLM outputs.

**Experiment-8:** How does the sentiment polarity of responses vary across different Large Language Models (LLMs)?

### Objective

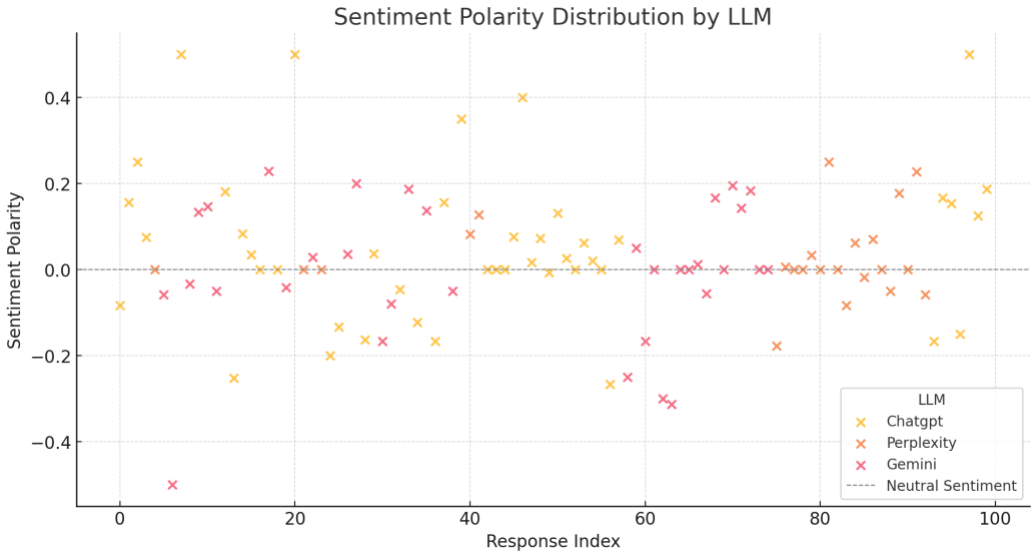
This experiment aims at finding out the impact of the five response categories of positive, negative or neutral sentiment polarity that the LLMs such as ChatGPT, Gemini, and Perplexity will display. This gives direction towards the tone and the sentiments of response that is given by these models.

## Dataset Preparation

1. **Data Source**:
    - The dataset includes:
        - Questions posed to the LLMs.
        - The answers produced by every single LLM.
        - Assessment of the valence score of each answer.
2. **Preprocessing Steps**:
    - **Sentiment Analysis**:
        - Use a sentiment analysis tool or library (e.g., TextBlob or VADER) to compute the sentiment polarity of each response.
        - Sentiment polarity ranges:
            - Positive: Polarity > 0.
            - Neutral: Polarity = 0.
            - Negative: Polarity < 0.
    - **Categorization**:
        - Match each response to the LLM with which it was completed so it can be further analyzed.
    - **Index Assignment**:
        - To simplify this a response index can be assigned:
3. **Columns Used**:
    - `Response by the top LLM`: : They were used to compute the sentiment polarity.
    - `Which top LLM you tried`: To group responses by LLM.

## Analysis

1. **Sentiment Distribution**:
    - Determine the percentage of positive, neutral and negative sentiment for each LLM.
    - Analyze whether certain LLMs lean towards specific sentiment types.
2. **Visualization**:
    - A sentiment analysis diagram is constructed as a scatter plot with ‘sentiment polarity ‘on the y-axis and ‘response index’ on the x-axis; each LLM type color coded.
3. **Comparative Analysis**:
    - Thus, it is required to compare the spread and range of the sentiment polarities across the different LLMs.




## Insights

1. **ChatGPT**:
    - Creates both positive, mixed, and even slightly negative attitudes.
    - Has a very similar distribution to the overall sentiment polarity, and displays less data dispersion at that.
2. **Gemini**:
    - Is somewhat closer to slightly positive sentiment polarities nonetheless, with rather significant negative extremes.
    - Responses show this range of polarity to suggest variability in the kind of tone associated with each.
3. **Perplexity**:
    - Has a possibility to come up with more balanced sentiment polarities and at the same time is characterized by more outliers both positively and negatively oriented.
4. **General Trends**:
    - Transitional responses are not rare across all LLMs, suggesting that models should achieve a middle tone.
    - Most words are positive, probably due to optimisation for easy to use basic positive words.

## Conclusion:

This particular project demonstrates not only the possibilities but also the flaws of large language models (LLMs) when it comes to solving intentionally flawed scientific questions intentionally designed to lead to incorrect answers. Filtering and adjusting questions from the ScienceQA Dataset and from project SciQ allowed revealing insightful findings about a particular kind of LLMs performance, including ChatGPT,

Gemini, and Perplexity. This reveals several concerns including Cognitive Limits of the models: They are logical fallacies, factual flaws, and question fallacies in terms of Physics, Biology and Mathematics areas of specialty. Perplexity and Gemini both outperformed ChatGPT on consciousness and creativity as a distinct aspect, although ChatGPT had areas of weakness in detail-oriented responses, which was also true for both Perplexity and Gemini.

This analysis highlights the need for fine-tuning at the domain level and the general need to incorporate better error tuning in LLMs. They also have to learn how to improve both the questions used and the context or background information in which the questions are asked and used so that the questions used become reliable for scientific use. In the wake of these limitations, we are able to lay down a roadmap for the enhancement of actual applications of LLM in education research and other fields so that these models can play an apt role in efficient reasoning and formulation of queries as per the need of a particular field.