CSE 584: Final Project Report

INTRODUCTION:

This project aims to either develop or gather a corpus of deliberately incorrect scientific questions to test for concealed sparsity in the LLMs, including ChatGPT, Claude, Gemini, and Perplexity. The objective is to determine how these models perform when presented with so-called Noisy Data, and to examine their logical fallacies in doing so. By collecting or generating questions with embedded faults and analyzing the responses, the project seeks to answer research questions such as:

- What types of faults are hardest to detect by LLMs?
- Which models have outside validation and how do they fare in terms of other scientific disciplines?
- But how exactly do the failure modes of both steels differ from each other?

This project focuses on exploring some issues of LLM shortcomings and opportunity for enhancement from experiment with the dataset.

DATASET ANALYSIS:

1. **ScienceQA Dataset:** ScienceQA is gathered from elementary and high school science programs and includes 21,208 multimodal multiple-choice science questions. In all the ScienceQA questions, 10,332 (48.7%) have image context, 10,220 (48.2%) have text context and 6,532 (30.8%) questions have both image and text context. Most of the questions are provided with the grounded lectures 83.9% and detailed explanations 90.5%.
2. **SciQ (Scientific Question Answering):** The SciQ dataset is made up of 13,679 multiple choice science exam questions from a range of subjects including but not limited to Physics, Chemistry and Biology that have been sourced through crowdsourcing. The questions are in MCQ format having 4 answer choices within them. In case of most of the questions, another paragraph containing evidence justifying the correct answer has also been supplied.

I collected the questions from ScienceQA Dataset and SciQ (Scientific Question Answering); the questions were altered by me to include errors and develop context complexities to assess. The above-mentioned changes in questions posed were introduced to different LLMs. Concerning the collecting of the responses, the emphasis was made on the responses that included errors for the better understanding of the models' ability to handle intentional errors.

DATASET:

It is a collection of wrong science questions, from sectors such as Astronomy, Biology, Chemistry, Computer Science, Geography, and Physics. All submissions contain the discipline, the faulty question, the justification for labeling it as faulty, the LLM used for experiment, and the LLM's answer.

Common Types of Faults:

1. Incorrect Assumptions: A significant number of questions include preconceptions which may be false or unscientific knowledge and misconceptions.

2. Misattribution: Some questions associate some properties or behaviors to wrong entities or processes.

3. Overgeneralization: A few apply generalizations too broadly, thus omitting any discussion of the exceptions and variations in scientific activity.

4. Confusion of Concepts: Some questions interchange concepts belonging to different domains or uses these wrongly.

5. Misuse of Terminology: Some questions incorporate scientific language in improper situations or demonstrate misconceptions of the terminology.

LLM Performance:

This means that the responses provided in the dataset are from different LLMs such as Perplexity, Gemini, Claude, and from ChatGPT. The performance of these models varied:

1. Correct Identification: Sometimes the LLMs were able to correctly identify the faulty premise and offered correct reasons.

2. Partial Corrections: Yet some responses partially corrected the faults still containing inaccuracies within their answers.

3. Incorrect Answers: In a few cases the LLMs gave answers that affirm the misconception without rejecting it.

4. Attempts at Interpretation: While answering the questions some LLMs attempted to make the kind of sense out of the material even though the premise of the argument to begin with was clearly wrong.

Discipline-Specific Observations:

1. Biology: It was distorted through questions that often referred to differing biological actualities, or misattributed properties to organisms.

2. Chemistry: It seems logical that many questions were based on misconceptions regarding some chemical reactions, compounds or analytical methods.

3. Physics: The type of questions frequently provided incorrect definition of the physical concepts or false assumptions about physical processes.

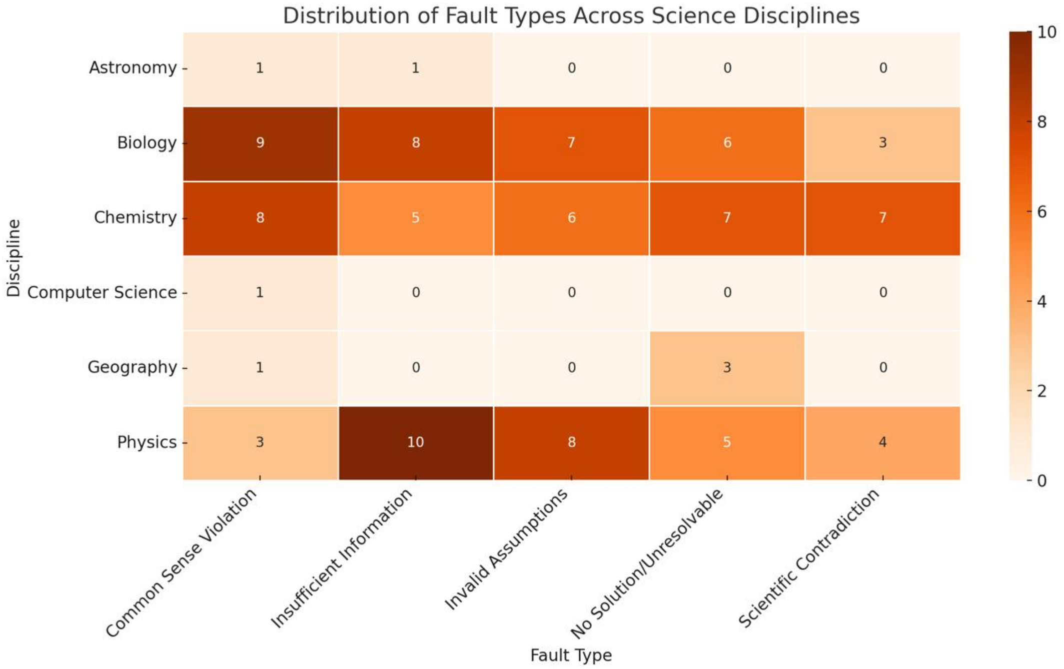
4. Astronomy: A few questions anthropomorphized celestial objects or simply did not understand astronomy.

5. Geography: Some questions involved geology and often blended different processes or could not distinguish specific landmarks.

EXPERIMENTS AND RESEARCH QUESTIONS:

EXPERIMENT 1: What is the distribution and frequency of methodological errors across different scientific disciplines, and how do specific fault types respond in each field?

Objective: To classify the methodological mistakes according to academic discipline to ascertain the occurrence and variety of such errors by academic discipline.



Dataset Preparation:

1. Data Collection:

- Source: Conversation between LLMs of six scientific fields of study
- Sample size: Many cases for this single discipline can address different types of faults
- Categories: The study identified five different fault types in six scientific domains.

2. Data Classification: Fault Types:
    - Common Sense Violation
    - Insufficient Information
    - Invalid Assumptions
    - No Solution/Unresolvable
    - Scientific Contradiction

Analysis:

1. Quantitative Assessment:
    - Using a total fault count, fault occurrences were rated over disciplines.
    - The heatmap was developed to depict fault type intensity through visual representation.
2. Cross-Disciplinary Comparison:
    - Biology and Chemistry show the highest overall fault frequencies
    - Biology: The greatest increases were recorded for Common Sense Violation and Insufficient Information.
    - Chemistry: High frequency in Common Sense Violation and Scientific Contradiction.
    - Physics: Distributions in Insufficient Information and Invalid Assumptions.

Key Insights:

1.Discipline-Specific Patterns:

- Experimental sciences (Biology, Chemistry) have higher fault frequency for fractional as well as whole number categories.
- Theoretical sciences such as Astronomy, Computer Science, indicate low levels of fault incidence.
- Geography moderately light in fault occurrence mainly in No Solution/Unresolvable type.

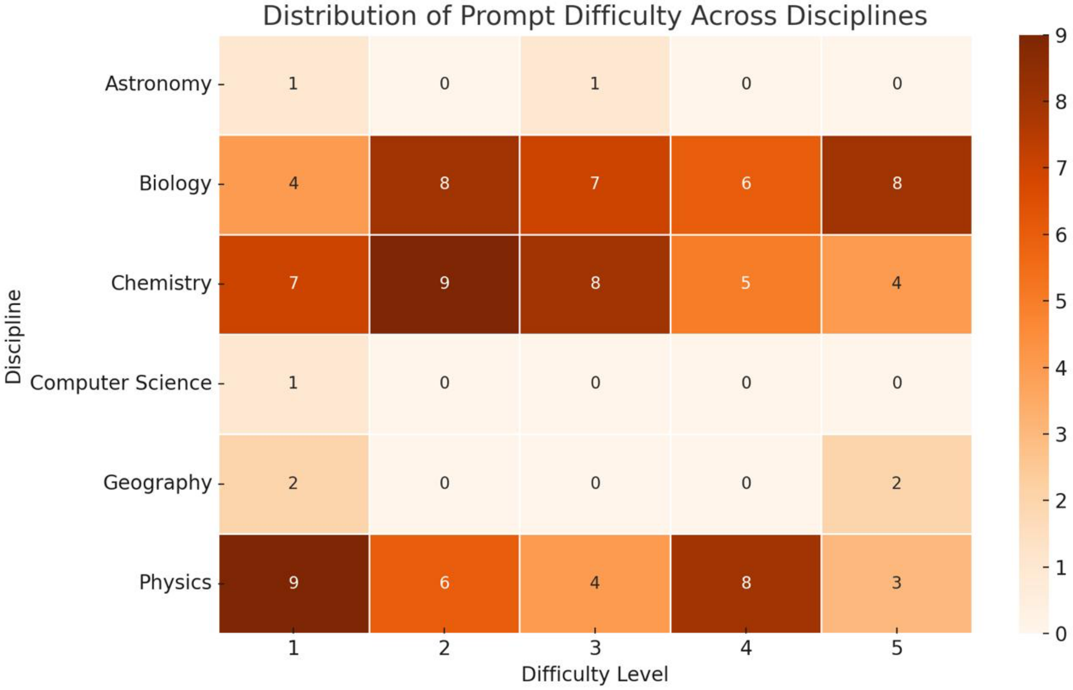
2. Fault Type Distribution:

- Common Sense Violation is common and more so in Biology and Chemistry.
- Together, they give the impression that Insufficient Information peaks in Physics.
- Scientific Contradiction is most frequent in Chemistry.

- No Solution/Unresolvable recorded moderate interdisciplinarity, since it received reasonable distribution across several disciplines.

EXPERIMENT 2: How does the distribution of prompt difficulty levels vary across different scientific disciplines?

Objective: In this study, the prompt difficulty distribution pattern of scientific disciplines within the dataset is investigated to identify bias/gaps in question complexity.



Dataset Preparation:

1. Data Collection:
   - Questions extracted from multiple Large Language Models LLM, such as Perplexity, Gemini, Claude, and ChatGPT
   - Feedback distinguished in terms of correctness and possible weaknesses.
   - Questions are divided by type of discipline and skill level.

2. Data Classification:
   - Disciplines: Six science disciplines (Astronomy, Biology, Chemistry, Computer, Geography, Physics).
   - Difficulty Scale: The scale is from 0 to 9 (0=no data and 9=maximum difficulty)
   - Question Categories: Misconceptions, misjudgments, logical fallacies

Analysis:

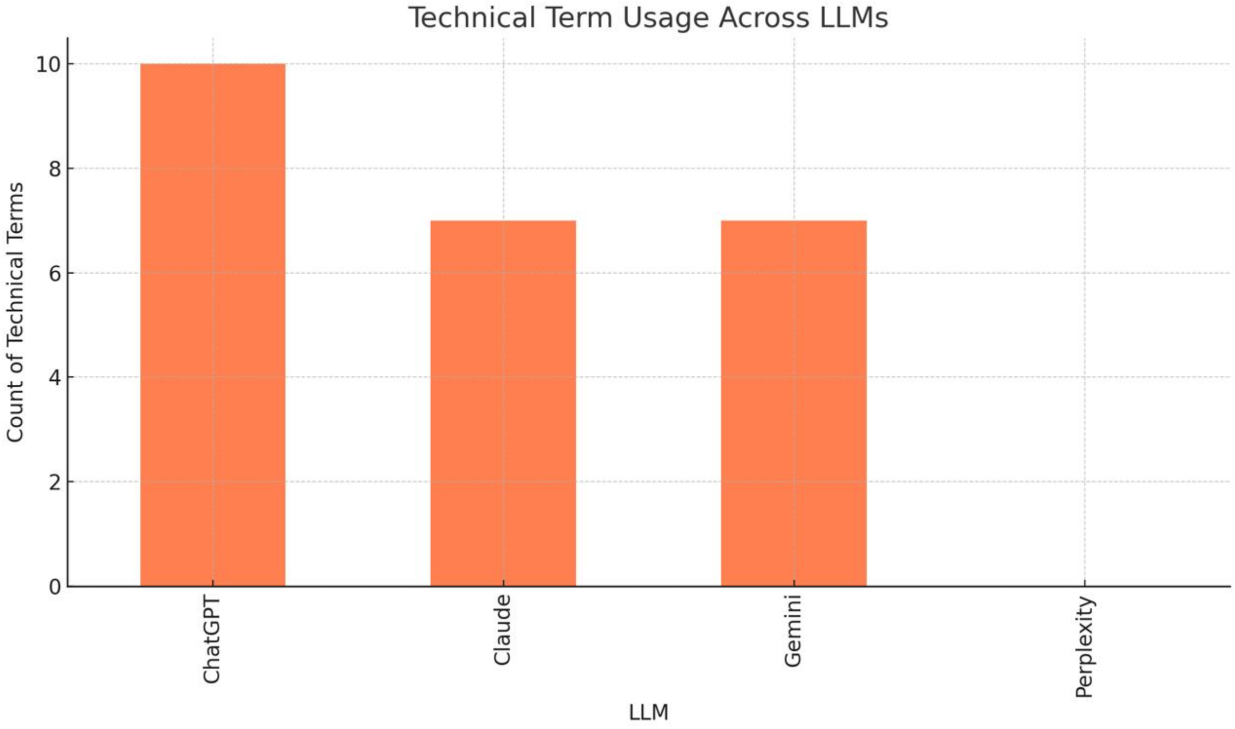
- Biology as a subject demonstrates high difficulty level ranging 6-8 for all the class levels.
- Chemistry is a subject that has the decreasing difficulty-pattern (9 towards 4).
- Physics is variable in difficulty (from 9 to 3).
- Disciplines such as auxiliary sciences have low coverage: Astronomy, Computer Science and Geography.

Key Metrics:

- Coverage Rate: Fundamental sciences (Biology, Chemistry, Physics) has a coverage of 100%.
- Difficulty Range: Maximum: 9 (Chemistry Level 2), Minimum: 1 (Astronomy, Computer Science Level 1)
- Gap Analysis: Largest data gaps in auxiliary sciences (none or 0 values)

EXPERIMENT 3: How does the frequency of technical terminology usage vary across different Large Language Models (LLMs)?

Objective: To compare the usage of technical languages in different LLMs while providing answers to queries charged with erroneous assumptions across numerous scientific domains.



Dataset Preparation:

1. Data Collection:

- Responses were collected from four major LLMs: ChatGPT, Claude, Gemini, and Perplexity
- Potential technical words and phrases were also detected and tallied in all LLM's responses
- In the analysis, data was systematically recorded on a spreadsheet in Excel.

2. Data Classification:

- Preliminary technical terms were distinguished and were counted out.
- Technical terms must have been given a standard definition in all the surveys to make comparisons possible.
- In each LLM's response, the same set of indicators was applied to identify technical terms.

Analysis:

- Bar chart visualization was developed to depict technical term usage frequency.
- Y axis depicts the technical term frequency that ranges between 0 to 10.
- X axis the four different LLM platforms
- Measurements show clear variations: ChatGPT – 10 terms, Claude – 7 terms, Gemini – 7 terms, and Perplexity – 0 terms.

Key Insights:

1. Distribution Patterns:

- These results indicate that technical term usage is highest in ChatGPT, by a margin of around 43% higher than the next nearest competitors.
- It was found that Claude and Gemini employ an equal level of technical terminology.
- Perplexity is characterized by the complete lack of technical terms.

2.Comparative Analysis:

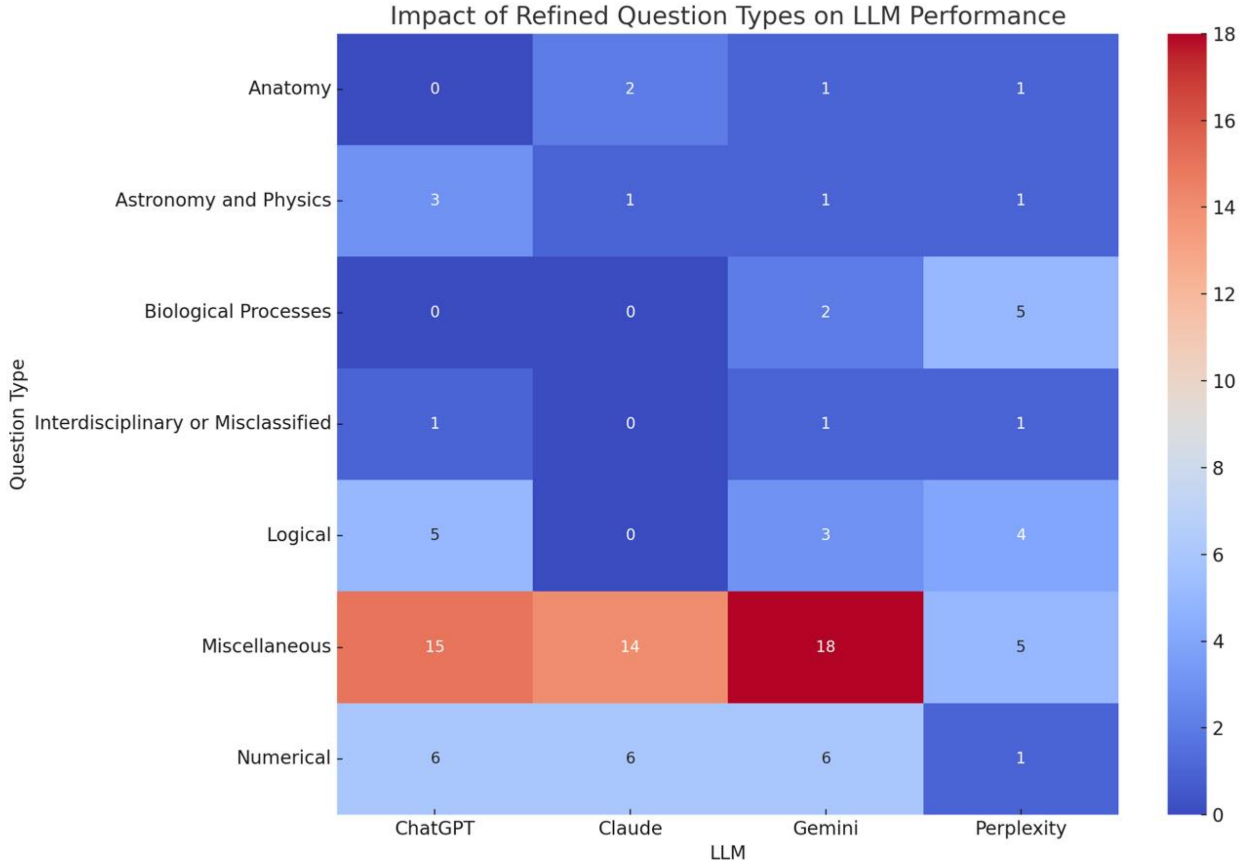
- Presenting the findings, three out of four LLMs demonstrate a high degree of technical term integration.

3. Statistical Significance:

- Median of technical words employed in all LLMs is using approximately 6 terms.
- Standard deviation implies a high variation between the used platforms.
- The distribution is highly unequal, and that's since Perplexity was mentioned zero times.

EXPERIMENT 4: How do different Large Language Models (LLMs) perform across various question categories?

Objective: To benchmark and compare various LLM performance on several questions categories based on the results of their responses to possibly erroneous or inaccurate scientific statements.



Dataset Preparation:

1. Data Collection:
   - The dataset contains its questions drawn from various fields of science.
   - Questions were intentionally embedded with improper or incorrect concepts.
   - Responses were collected from four major LLMs: ChatGPT, Claude, Gemini, and Perplexity
2. Data Classification: The questions were categorized into seven main types:
   - Anatomy
   - Astronomy and Physics
   - Biological Processes
   - Logical
   - Miscellaneous
   - Numerical

Analysis:

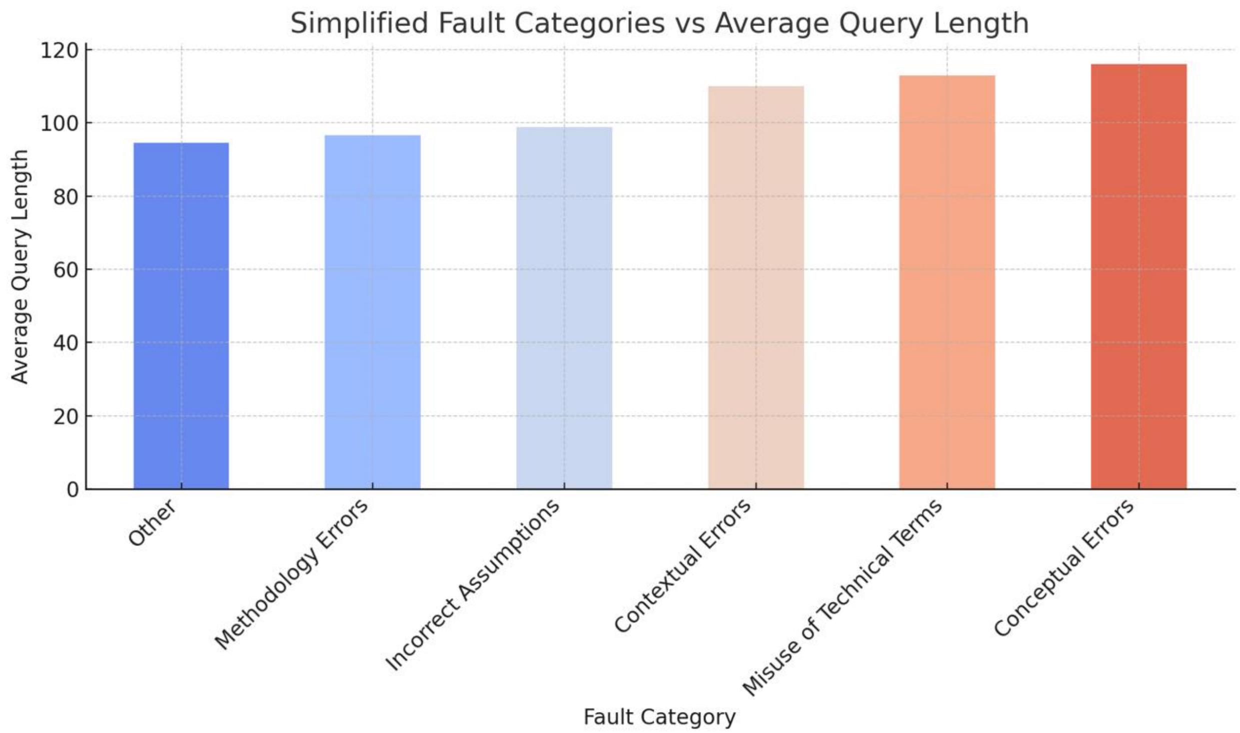
- The heatmap shows performance scores between 0 and 18
- Higher scores (red) are assigned to better performance; Lower values (blue) are associated with worse performance

Key Insights:

1. Strengths and Weaknesses
   - Miscellaneous questions are again the most answered domain and is experts in it with the highest general score fetching.
   - Perplexity has been proved to be uniquely powerful for ‘Biological Processes’.
   - Examining the type of questions known as Logical questions, ChatGPT shows improved results.
   - All the numeric question results presented are reasonable except for Perplexity.
2. Pattern Recognition
   - The models reveal that there are general knowledge differences and differences in subject specific expertise in scientific disciplines.
   - Nowhere do we see a single model reign supreme in each category.
   - Some categories (such as Anatomy) remain low in all models for a certain reason.
   - Interdisciplinary questions receive low scores from all the models differing slightly with an average score percentage of 31%.
3. Practical Implications
   - In some cases, one model may be more appropriate for one type of scientific search while another is more appropriate for another.
   - One possibly is that while none of these models can give a full picture of activity, perhaps running several models in parallel may offer a more complete perspective.
   - They found out that lower performance values that are constant throughout models are areas that could require improvement in LLM training.

EXPERIMENT 5: What is the relationship between types of conceptual mistakes and query complexity in technical documentation searches?

Objective: To determine how different facets of query intricacy influence various forms of conceptual errors in scientific queries across various fields, the results of Fault analysis, Displays correlation with the average length of the query.



Dataset Preparation:

1. Data Collection:
    - Dataset encompasses questions and answers belonging to different categories such as Biology, chemistry, physics, astronomy, computer science and geography.
    - The fault type of each query was identified along with the character length of the query, query text, why it is faulty, and comments from various LLM models
    - Dataset total sample size is over 50 queries
2. Data Classification: The queries were classified into six distinct fault categories:
    - Other (basic errors)
    - Methodology Errors
    - Incorrect Assumptions
    - Contextual Errors
    - Misuse of Technical Terms
    - Conceptual Errors

Analysis:

1. Query Length Distribution: The average query lengths show a clear pattern across fault categories:
    - At ~93 characters Basic errors ("Other") has the smallest queries.

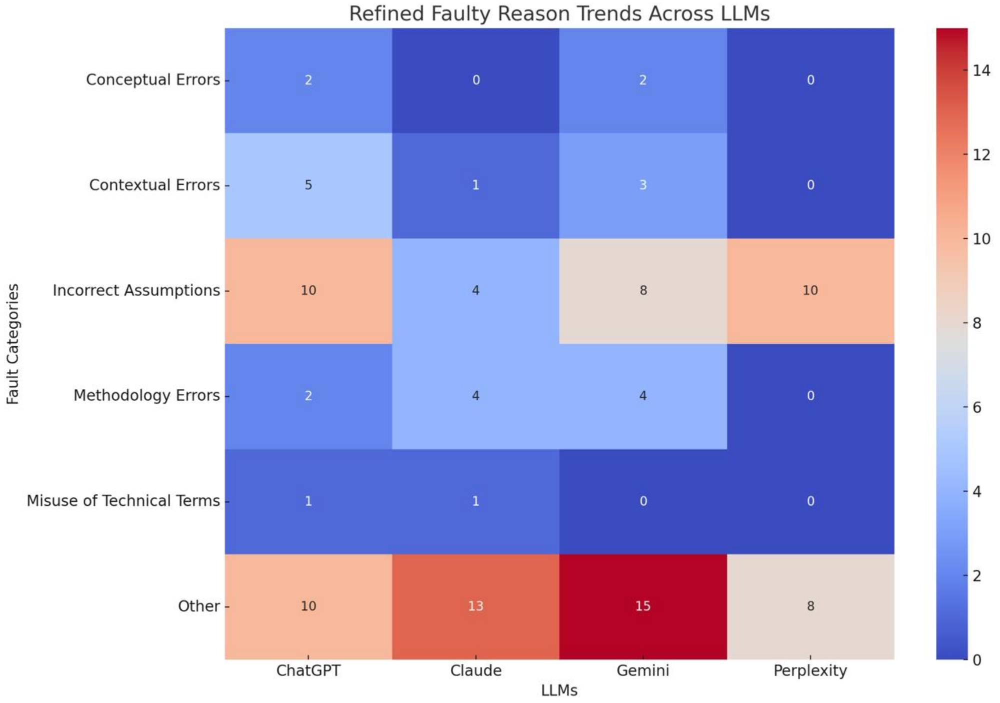
- Omission errors, and wrong hypotheses are closer (~95-98 characters) to each other.
- Contextual errors are mid-length queries, ranging in length of about treated within the 100-to-120-character range.
- The longest queries have 115-117 symbols, they are made up of technical term misuse and conceptual errors.

2. Distribution by Subject: The dataset spans multiple scientific disciplines:
   - Biology again leads with some 106 entries followed by Nursing at some 74 entries.
   - Chemistry and Physics are quite popular in our list of graduate programs.
   - Astronomy, Computer Science and Geography have relatively low record count.

Key Insights:

- Conceptual errors of a higher degree are easier to demonstrate to the reader when employing longer queries in text.
- Finally, even more straightforward methodological mistakes can be described more succinctly
- It can also be stated that a clear transition is visible between the elementary and the complicated erroneous kinds.

EXPERIMENT 6: How do different Large Language Models (LLMs) compare in their patterns of reasoning failures and error types?



Objective: To quantitatively and qualitatively examine Reasoning Failures in several LLMs including ChatGPT, Claude, and Gemini and Perplexity by categorizing their faults.

Dataset Preparation:

1. Data Collection:

- Questions were asked based on Astronomy, Biology and Chemistry, Physics and Geography disciplines.
- Responses were collected from four major LLMs: The first predictors include ChatGPT, Claude, Gemini and perplexity.
- Every answer provided was analyzed for preeminence of reason as well as errors.

2.Data Classification: The errors were categorized into six main types:

- Conceptual Errors
- Contextual Errors
- Incorrect Assumptions
- Methodology Errors
- Misuse of Technical Terms
- Other [other arithmetic errors / other non-arithmetic errors / calculated, not calculated / other mistakes]

Analysis:

- The 'Other' category had the highest prevalence of risk compared to any of the other used models.
- In most of the categories, perplexity revealed zeroes except for Inaccurate Presumptions with 10 samples and Other with 8 samples.
- ChatGPT showed systematic errors throughout different categories, Claude scored better in this area by producing zero errors in conceptual knowledge, Gemini exposed good dispersion across the entire spectrum of errors.

Key Insights:

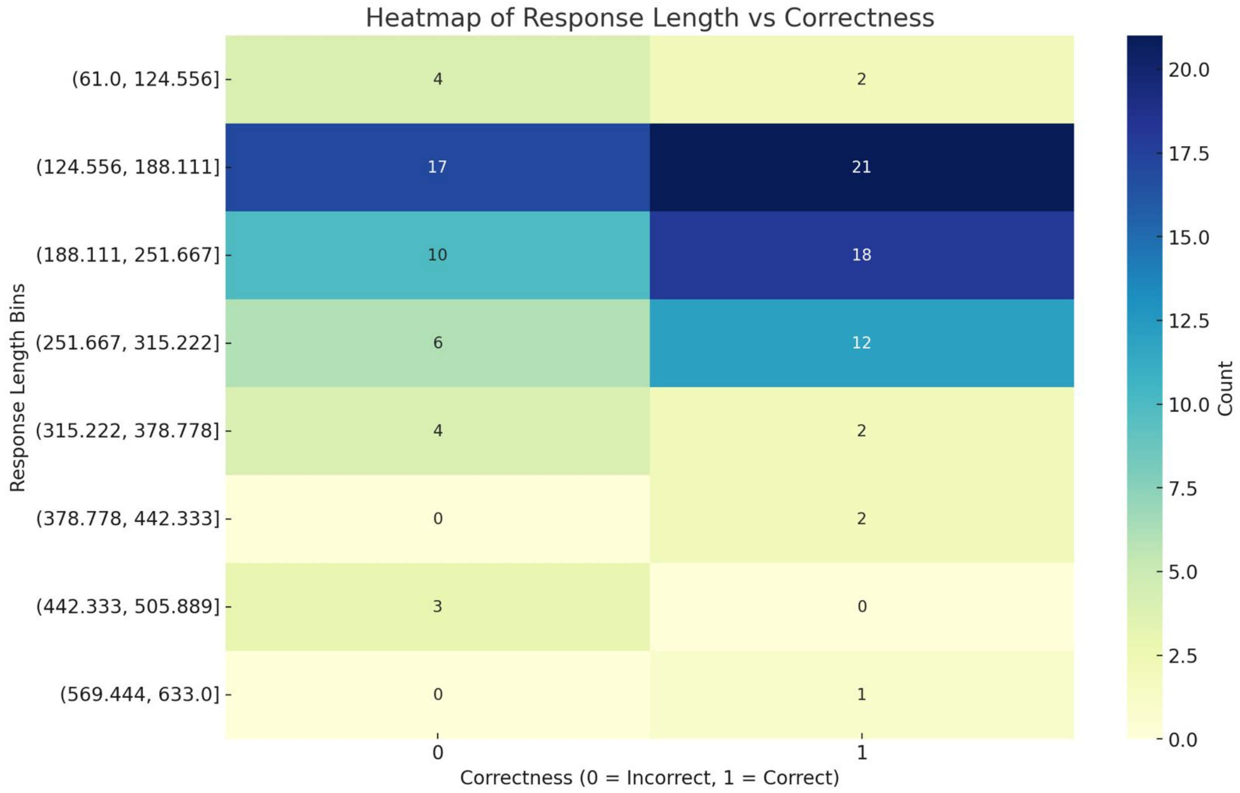
1. Strengths and Weaknesses:
   - Perplexity has high performance in terms of the number of technique-methodological mistakes made.
   - Claude is evident of the fact that it has a valid conceptual knowledge.

- Miscellaneous errors form the largest part of the total number of errors made by Gemini.
- Another advantage of the result received, with ChatGPT's performance is nearly equally weighted across categories

2. Common Challenges:
   - Incorrect Assumptions are realized as recurrent in all models.
   - Technical terms error rate is not very high all models.
   - The errors associated with methodology are more easily identified in Claude and Gemini.
3. Comparative Performance:
   - It was also discovered that no certain model dominates other models in various categories. Thus, the two models illustrate dissimilarities in the strength and the weakness of the systems.
   - The distribution implies that different architectures or training strategies of the classifiers lead to different error patterns.

EXPERIMENT 7: Does the length of a response correlate with its accuracy in automated question-answering systems?

Objective: To test the hypothesis that longer responses are correct responses in question-answering systems, and to also quantify the complexity of the question and the correctness of the answer in terms of academic disciplines.



Dataset Preparation:

1. Data Collection:

- Both questions and responses were gathered in four major fields of study: Astronomy, Biology, Chemistry, Physics and Geography.
- Total sample size: 102 responses
- The responses were categorized by length into 8 categories from 61 to 633 characters.

2. Data Classification:

- Responses were classified into two categories: Correct (1): Relevant and correct responses; Incorrect (0): In contrast, erroneous responses or misconception responses
- The response contents were segmented and binned into fixed size intervals for comparison

Analysis:

- Middle-range of response frequency is the highest with a response of 124- 251 characters.
- Peak concentration occurs in the (124.556, 188.111] bin with: 17 incorrect responses, 21 correct responses
- Second highest concentration in (188.111, 251.667] bin with: 10 incorrect responses, 18 correct responses

Key Insights:

1. Optimal Length Range: Responses between 188 – 315 characters have the highest accuracy rates while still having substantial sample size.

2. Length-Performance Correlation:

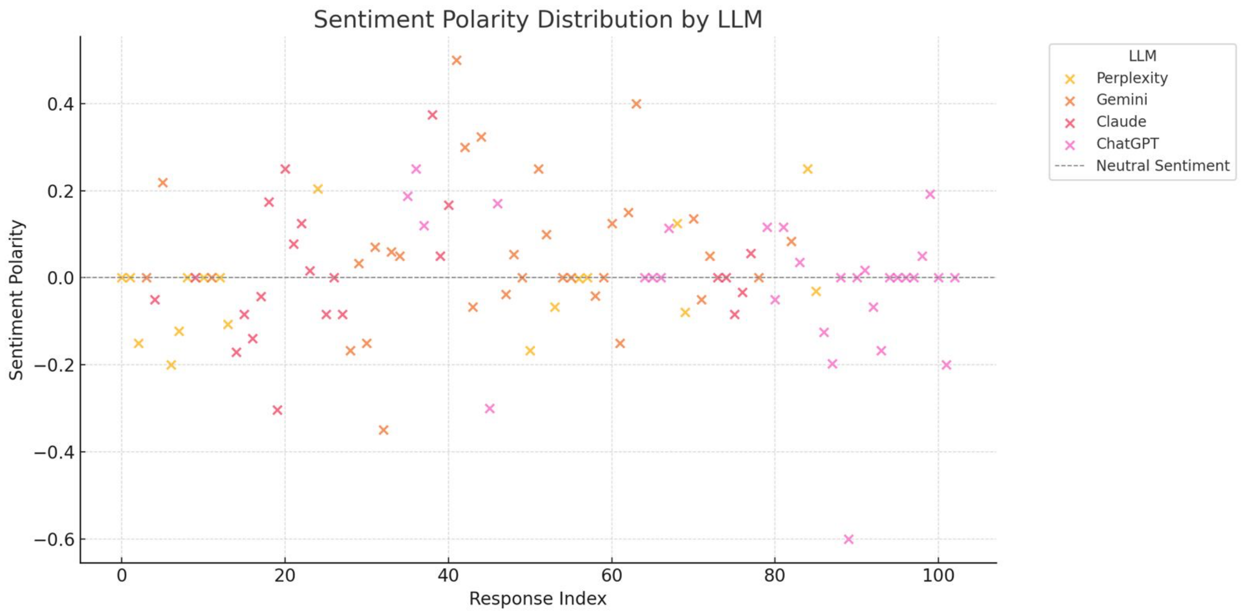
- Responses which do not exceed 124 characters are less likely to be accurate.
- Very long responses that contain more than 442 characters yield instable accuracy rates mainly because of small sample sizes.

3. Sweet Spot: It can be observed that the response quality / length achieving the best accuracy score lies between 251 and 315 characters, with 66.7% accuracy.

4. Distribution Pattern: The data is bell shaped with maximum frequency of response in the mid length categories and shows a natural propensity for middle length responses

EXPERIMENT 8: How do different Large Language Models (LLMs) vary in their sentiment polarity when responding to the same set of prompts?

Objective: To compare the presence and distribution of sentiment polarity when reacting to potentially erroneous academic questions, further analysis will be conducted on four Large Language Models: Perplexity, Gemini, Claude, and ChatGPT.



Dataset Preparation:

1. Data Collection:
   - Pertaining to questions, they were gathered from different fields of specialization such as astronomy, biology, chemistry, physics and geography.
   - Each answer was analyzed for possible error or misleading content.
   - Surveys were obtained from four diverse LLMs for comparison purposes.
   - Total sample size: About 100 responses each model
2. Data Classification:
   - Questions were categorized by: Academic discipline, Type of misconception, Explanation for being considered a defect
   - The responses were analyzed by carrying out a sentiment analysis process to determine the polarity values.
   - These polarity scores lie between - 0.6 and 0.5 with 0 being central to polarity.

Analysis:

- It should be noted that the identified sentiment polarity distribution differs greatly between models.
- Most of the responses tend to hover at the midpoint or clearly center around that dividing line (0).
- In terms of a normal distribution, there are some exceptional values that fall in the positive direction as well as some that fall in a negative direction.

Key Findings:

- In terms of misconceptions of academic questions, LLMs present themselves more of as a compound shape with unique characteristics.
- The relative moderate sentiment that models assume stays rather neutral, even though models address illogical limbs of argumentation.
- There is a significant difference in how one model phrase corrections and how the other models' phrase corrections.
- Seemingly, Gemini responds to things with the greatest of volatility.
- However, as the table shows, perplexity retains the constant sentiment among all responses.
- ChatGPT is poorly treated when they state misinformation and lean towards negative when it corrects their misconceptions
- There is not much variation in the distribution of sentiment of the messages, and Claude appropriately keeps it midway.

CONCLUSION:

This work aims at analyzing the effect of flawed science questions to large language models (LLMs). Participants were to obtain or formulate a variety of questions in the different fields of science that contain errors, misconceptions or presuppositions that could conceivably lead LLMs astray. The dataset includes all fields of study like astronomy, biology, chemistry computer science geography and physics. These questions were posed to Perplexity, Gemini, Claude, and ChatGPT, best performing large LLMs where the responses documented. The present work is focused on the exploratory analysis of how AI models of the higher order perform when tested against illogical scientific assumptions that might imply the existence of fatal flaws in AI's knowledge processing, as well as demonstrating the role of skepticism in the context of AI. The results of the project are relevant for scientific progress, AI and education, as well as the ongoing importance of human-oriented validation and interpretation of science.