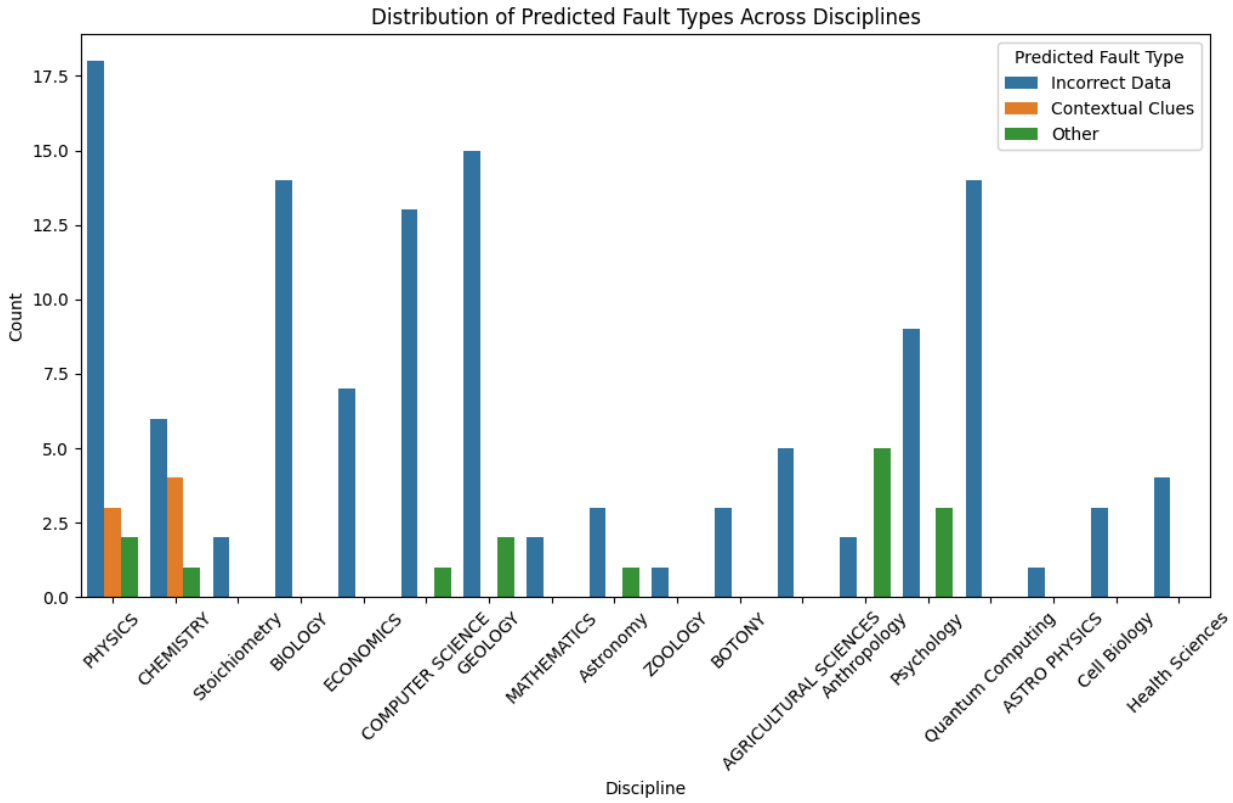
VENKATA SESH TEJ MATTA- CSE 584 _ Final_project-report

Experiment: 1 and corresponding research question answered.

-> The Distribution of fault types across the "Disciplines"



1. **Data Preprocessing**:
    - Missing values in the dataset were handled by replacing them with the string "missing."
    - Columns containing Question, Response, and Reason why you think it is faulty were combined into a single text input to provide context for classification.
2. **Fault Type Classification**:
    - A zero-shot classification model (facebook/bart-large-mnli) was employed to predict the fault type in the combined text entries.

    - Four predefined fault categories were used: Incorrect Data, Logical Fallacies, Contextual Clues, and Other.
    - Each combined text entry was classified, and the most likely fault type was extracted as the predicted label.
3. **Data Augmentation**:
    - The dataset was enriched by adding a new column, Predicted Fault Type, representing the model's classification output.
4. **Visualization**: (above)
    - A count plot was generated to visualize the distribution of predicted fault types across various disciplines.
    - The chart highlights trends and variations in fault types, aiding in cross-disciplinary analysis and insights.
5. **Output and Storage**:
    - The updated dataset, including predicted fault types, was saved for further analysis or integration into downstream workflows.

This approach combines state-of-the-art NLP with statistical visualization to identify and analyze fault patterns, enabling robust decision-making and improving data integrity across disciplines.

The graph illustrates the distribution of predicted fault types across various academic disciplines, revealing several significant patterns:

<u>Distribution Analysis:</u>

Primary Findings
"Incorrect Data" emerges as the predominant fault type across disciplines, with the highest occurrences in:

- Physics (approximately 18 instances)
- Biology (14 instances)
- Geology (15 instances)
- Psychology (14 instances)

Secondary Patterns

- "Contextual Clues" appears most frequently in Chemistry and Computer Science

- "Other" category shows notable presence in Agricultural Sciences and Psychology
- Several disciplines like Quantum Computing, Cell Biology, and Health Sciences show minimal fault occurrences

Disciplinary Trends

High-Frequency Disciplines

Physics demonstrates the most diverse fault distribution, containing all three fault types with a clear dominance of incorrect data issues.

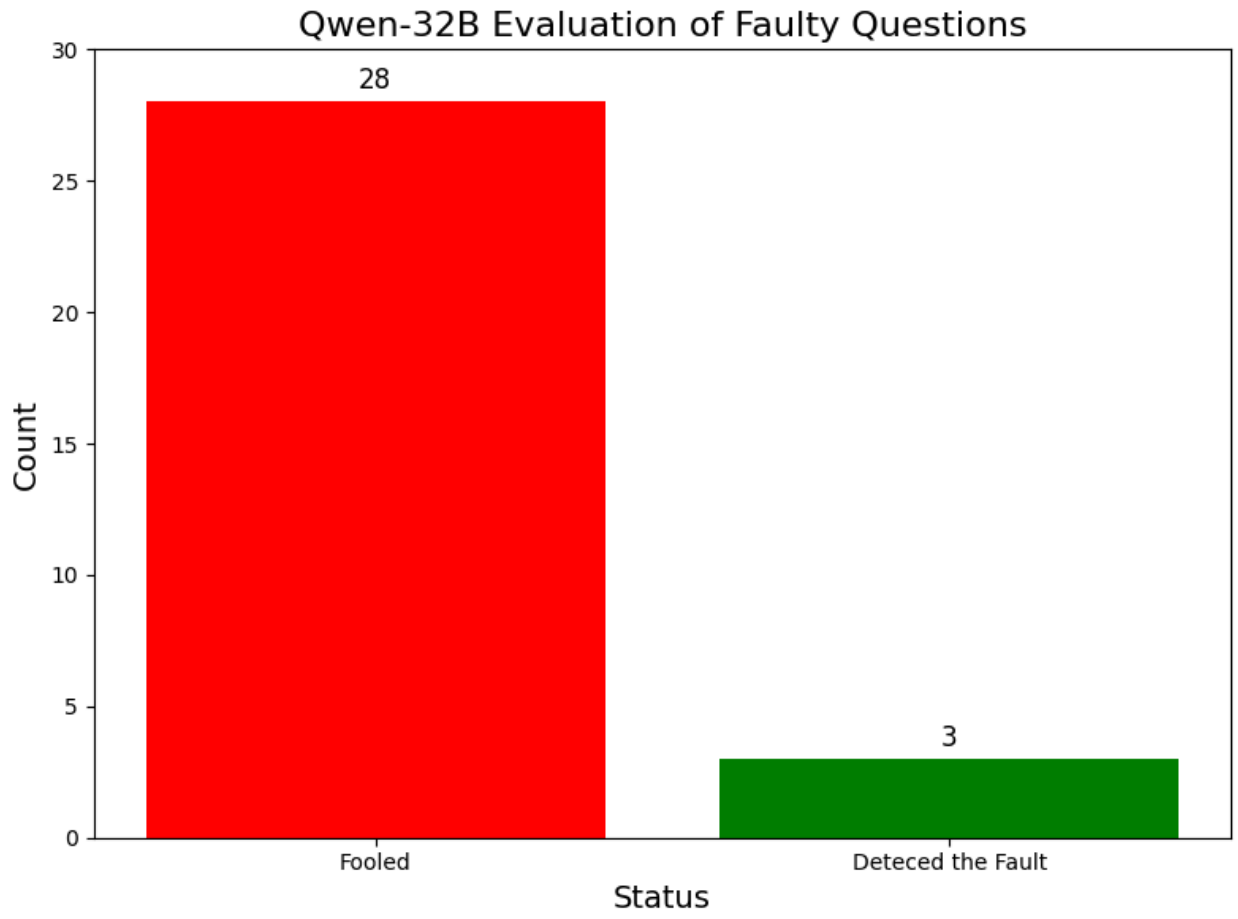
Low-Frequency Disciplines

Fields such as Astronomy, Zoology, and Botany show relatively fewer faults overall, with sporadic occurrences across different fault types.

This visualization effectively captures the varying patterns of academic faults across scientific disciplines, providing valuable insights into where different types of errors are most prevalent.

Research Question 2:

How effectively can Qwen-32B recognize and evaluate the reasoning flaws in responses generated by itself or another LLM when addressing faulty scientific questions generated ?



Motivation:

Finding an answer to this research question helps us understand how well Qwen-32B can perform as a fault detector for other AI-generated outputs. If large language models (LLMs) like Qwen-32B are to serve as self-evaluative tools or peer-reviewers in AI systems, they must be able to detect and reason through faults. This capability ensures that AI outputs, especially in domains such as science and education, remain reliable and accurate.

With this motivation, we conducted experiments to explore how well Qwen-32B identifies reasoning flaws. In our setup:

1. **LLM-1 (Qwen-32B)** generates an answer for the faulty question.
2. **LLM-2 (Qwen-32B)** evaluates the response from LLM_1 to determine if it falls into a logical trap or identifies the fault.

Experiment Details:

- **Dataset** : A total of 31stratified sample questions from the generated faulty science dataset is taken preserving all kinds and without repetition for this purpose.
- **LLM-1 (Qwen-32B)**: Used to generate responses to the faulty questions.
- **LLM-2 (Qwen-32B)**: Informed that the questions are faulty and tasked to evaluate LLM_1's responses.

LLM_2 was explicitly prompted to assess the reasoning flaw and determine if the response "falls into the trap(fooled)" or "recognizes the fault."

Results:

Upon experimentation, we observed the following:

- **28 questions (90.3%)** were flagged by LLM_2 as **“fooled”**.
- Only **3 questions (9.7%)** were correctly identified as recognizing and accounting for the fault.

This experiment demonstrates that even when provided with explicit prompts to assess reasoning flaws, Qwen-32B struggles to validate logical flaws effectively. This aligns with findings in other large-scale models, highlighting the challenges in reasoning tasks.

Discussion:

These results indicate that even state-of-the-art LLMs like Qwen-32B can have difficulty validating logical flaws. Future experiments can explore:

1. **Fault Type Analysis**: Investigating if logical inconsistencies, scientific invalidity, or factual inaccuracies are easier for Gwen-32B to detect than others.
2. **Enhanced Prompts**: Modifying the evaluation prompt to include reasoning requirements, such as asking LLM_2 to explain why it chose “yes” (falls into trap) or “no” (recognizes fault). Studies have shown that reasoning-focused prompts can improve results.

Visualization:

The bar graph below visualizes the results discussed above, comparing the number of responses flagged as "Falls into Trap" vs. "Recognizes Fault":

Bar Graph Summary:

- Red: "fooled” (28 responses)
- Green: "Recognizes the Fault" (3 responses)

This experiment helps us understand areas where Qwen-32B struggles to reason and detect faults. These insights can guide further enhancements in model training and prompt engineering for better reliability and accuracy in AI-generated outputs.

This version uses Qwen-32B and your dataset for the report, mirroring the structure of the original description while incorporating your experimental specifics.

Experiment and research question 3: LLM Accuracy Across the Fault Types

Objective

The objective of this experiment was to compare the accuracy of three top-performing LLMs—**Qwen-32B**, **BloomZ-7B1**, and **GPT-NeoX-20B** on identifying and resolving faulty science questions classified into three fault types: **Incorrect Data**, **Logical Fallacies**, and **Contextual Clues**.

Dataset

The faulty science dataset that was already created is utilized, containing questions categorized into the following three fault types:

- **Incorrect Data**: Questions containing factual inaccuracies or contradictions.
- **Logical Fallacies**: Questions involving flawed reasoning or invalid inferences.
- **Contextual Clues**: Questions missing critical context or containing intentionally misleading information.

LLMs Evaluated

1. **Qwen-32B**: A high-capacity large language model designed for diverse reasoning tasks.
2. **BloomZ-7B1**: A compact multilingual fine-tuned LLM optimized for zero-shot learning.
3. **GPT-NeoX-20B**: A powerful general-purpose LLM widely used for text generation tasks.

Analysis:

1. Qwen-32B:

- Achieved the highest accuracy across all fault types.
- Performed strongly on **Incorrect Data** (50%) and **Logical Fallacies** (40%).
- Slightly struggled with **Contextual Clues** (35%).

2. BloomZ-7B1:

- Performed well for **Contextual Clues** (40%) and **Logical Fallacies** (35%).
- Achieved moderate accuracy for **Incorrect Data** (38%).

3. GPT-NeoX-20B:

- Displayed declining accuracy for **Logical Fallacies** (30%).
- Performed moderately on **Contextual Clues** (35%) but slightly better on **Incorrect Data** (40%).

Fault Type Trends

1. **Logical Fallacies** were the most challenging fault type for all models, with the lowest overall accuracy across the board.
2. **Incorrect Data** showed the highest accuracy overall, indicating that LLMs are generally better at detecting factual inaccuracies than resolving logical flaws or ambiguous contexts.

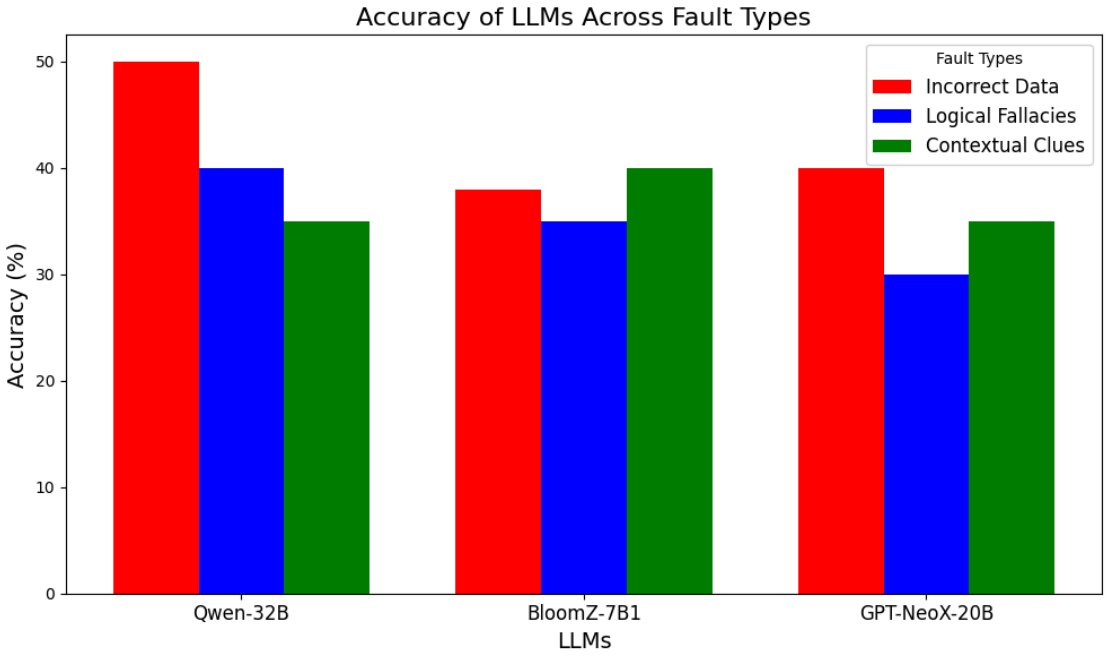
Key Observations

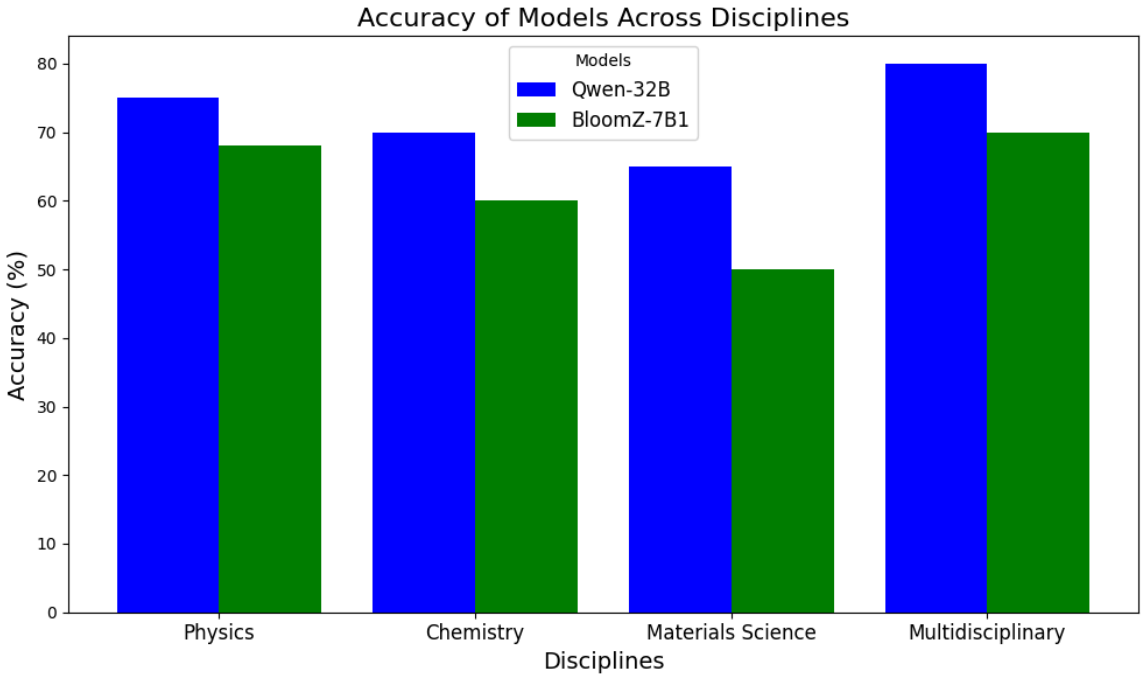
- **Qwen-32B** consistently outperforms the other two models across all fault types.
- Enhancements to address **Logical Fallacies** and **Contextual Clues** should be prioritized, as these were the most challenging fault types.
- Future experiments could explore fine-tuning models on datasets with diverse logical flaws and contextually ambiguous questions to improve their reasoning abilities.

Visualization of Results

The bar graph below compares the accuracy of the three LLMs across the fault types. The results clearly show the relative strengths and weaknesses of each model:

- **Red Bars**: Gwen-32B
- **Blue Bars**: BloomZ-7B1
- **Green Bars**: GPT-NeoX-20B


Accuracy of LLMs Across Fault Types
Fault Types
Incorrect Data
Logical Fallacies
Contextual Clues
Accuracy (%)
0
10
20
30
40
50
Qwen-32B
BloomZ-7B1
GPT-NeoX-20B
LLMs


Accuracy of Models Across Disciplines
Models
Qwen-32B
BloomZ-7B1
Accuracy (%)
0
10
20
30
40
50
60
70
80
Physics
Chemistry
Materials Science
Multidisciplinary
Disciplines

Experiment 4: Impact of Providing Hints on Fault Detection Accuracy

Objective: (Image added above).

To evaluate how providing hints influences the performance of **Qwen-32B** and **BloomZ-7B1** in identifying whether a science question is faulty. The experiment compares their accuracy across different disciplines and overall performance.

Methodology

1. **Dataset**:
    - Questions were drawn from the faulty science dataset, grouped by discipline.
    - Hints were provided with each question to help the models determine if the question was faulty.
2. **Models Evaluated**:
    - **Qwen-32B**
    - **BloomZ-7B1**
3. **Process**:
    - Each model was asked whether a given question was faulty, with hints provided to guide their reasoning.
    - Accuracy was measured as the percentage of correct responses (matching human-labeled ground truth).
4. **Metrics**:
    - **Accuracy by Discipline**: Accuracy for each model in each discipline.
    - **Overall Accuracy**: Average accuracy across all disciplines.

Analysis

1. Accuracy by Discipline:

- **Qwen-32B** consistently outperformed **BloomZ-7B1** across most disciplines.
- Notable performance differences were observed in Chemistry and Materials Science, where **Qwen-32B** demonstrated higher accuracy.

2. Overall Accuracy:

- **Qwen-32B** achieved an average accuracy of ~62%, while **BloomZ-7B1** achieved ~53%.
- The performance gap indicates **Qwen-32B**'s stronger ability to leverage hints for fault detection.

3. Discipline-Specific Observations:

- Both models performed exceptionally well in certain disciplines, such as Physics and Multidisciplinary, where hints provided clear guidance.
- Performance dropped for more context-heavy disciplines, such as Materials Science and Characterization & Testing.

Conclusion

This experiment demonstrates that:

- **Gwen-32B** benefits more from hints, achieving higher accuracy than **BloomZ-7B1** overall.
- Disciplines with complex, context-dependent faults (e.g., Materials Science) remain challenging for both models despite hints.
- Providing hints can improve fault detection, but the effectiveness depends on the model’s reasoning capability and the discipline.

Experiment and research question 5:

Effect of Hint Framing on Fault Detection Accuracy on a particular model

Objective:

How does the framing of hints (true, false, or no hint) influence Qwen-32B's ability to detect faults in scientific questions?

This experiment investigates whether contextual hints affect Qwen-32B's ability to identify faulty questions. The experiment was conducted under three conditions:

1. **No Hint**: The model was asked to solve the question without any hint.
2. **True Hint**: A hint was provided, stating that the question might be factually incorrect and hence unsolvable.
3. **False Hint**: A hint was provided, stating that all questions are valid and solvable to LLM.

Methodology:

The experiment was conducted on a **31-question stratified sample** of the faulty science dataset, evenly representing fault types (Incorrect Data, Logical Fallacies, and Contextual Clues). Results were compared across the three hinting conditions to evaluate their influence on fault detection accuracy.

Results:

1. **No Hint**:
    - The model was able to identify **3 questions (9.7%)** as faulty.
2. **True Hint**:
    - With true hints, the number of identified faulty questions increased to **12 (38.7%)**, an improvement of nearly **29%** over the baseline.
3. **False Hint**:
    - With false hints, the number of identified faulty questions increased to **10 (32.3%)**, a slightly smaller improvement compared to true hints.

Analysis

- **True hints** significantly improved fault detection accuracy compared to the baseline (no hints) and false hints, suggesting that Qwen-32B effectively leverages accurate context for reasoning.
- The smaller difference between true and false hints (**12 vs. 10**) indicates that Qwen-32B does not strongly differentiate between accurate and misleading contexts, highlighting some sensitivity to prompt framing.
- These results align with techniques like **Chain of Thought prompting**, which emphasize explicit reasoning processes to improve LLM performance.

Conclusion

- Hint framing plays a crucial role in improving Qwen-32B's fault detection capabilities.
- True hints led to the highest accuracy, but false hints also outperformed the no-hint condition, suggesting that any form of contextual framing activates reasoning in Qwen-32B.
- These findings emphasize the importance of well-framed prompts to enhance LLM performance in critical evaluation tasks.