Research Questions and Experiment

The emergence of advanced large language models (LLMs) has revolutionized problem-solving capabilities across various domains, including mathematics, science, and general knowledge. However, these systems remain susceptible to subtle errors, including falling for misleading or faulty questions that require reasoning beyond surface-level understanding. Identifying and addressing these vulnerabilities is critical for enhancing the robustness of LLMs, particularly in high-stakes applications such as education, scientific research, and decision-making.

This study builds on prior work, such as *FaultyMath*, which explored the creation of faulty math problems to challenge and evaluate LLM reasoning. Inspired by the methodologies in *FaultyMath*, I extended this approach to develop a diverse dataset of faulty science questions. By leveraging prominent datasets like **ScienceQA** and **SCIQAG**, I curated a specialized dataset aimed at testing the resilience of state-of-the-art LLMs in reasoning through intentionally misleading science problems.

Our objective is twofold: (1) to create a dataset of faulty science questions across diverse categories, ensuring coverage of various scientific domains and problem structures, and (2) to design and conduct experiments to evaluate how well leading LLMs, such as Gemini 1.5 Flash, GPT-4, and others, can detect and address these faulty problems. This paper describes the curation process for the dataset, the experimental framework, and preliminary results, providing a foundation for further exploration into LLM reasoning limitations.

Dataset Description

The dataset curation process closely mirrors the methodology used in *FaultyMath*, but it was adapted to address challenges specific to scientific reasoning. The key steps in the dataset creation pipeline are described below.

Source Datasets

To ensure diversity and relevance, I used **ScienceQA** and **SCIQAG** as the source datasets for valid science questions. These datasets provide a robust foundation, containing questions that span multiple scientific disciplines, including physics, biology, chemistry, and earth sciences. Both datasets include questions designed for varying levels of difficulty, making them suitable for adaptation into faulty versions.

Faulty Science Question Generation

To generate faulty versions of valid science questions, I utilized open-source LLMs, including **Qwen 14B**, **Qwen 32B**, **Llama 3.1 (3B)**, and the experimental **QwQ** model developed by the Qwen Team. **QwQ - 32B Preview** is a open source competitor to **GPTo1 - Preview. It even beat o1 in many benchmarks and it is 32B only.**

The process involved the following steps:

1. **Generation**: QwQ was used to transform valid questions from the source datasets into faulty science questions. Each valid question was paired with its correct answer and context, and QwQ was prompted to:
    - Alter the question in a way that introduces ambiguity, inconsistency, or false information.
    - Provide a brief explanation of why the resulting question is faulty.
2. **Verification**: The generated faulty questions were reviewed using **Qwen 32B**, which performed a self-verification process similar to that described in *FaultyMath*. The verification included:
    - Solving the faulty question.
    - Explaining why the question is faulty.
    - Assigning a classification label (A, B, or C) based on the difficulty and validity of the faulty question.

Self-Filtering Process

The self-verification process employed by Qwen 32B was critical in refining the dataset. The classification labels were interpreted as follows:

- **A**: The question is a faulty science problem but not challenging to identify as such.
- **B**: The question is a genuine faulty science problem that is challenging to identify.
- **C**: The question is valid and solvable, despite being labeled as faulty initially.

Only questions classified as **B** were retained, as they represent genuinely faulty and challenging problems. Questions labeled **A** or **C** were filtered out to ensure the dataset's quality and focus. Of the initial 5,000 questions generated, approximately 200 high-quality faulty science questions were selected for inclusion in the final dataset.

Testing and Validation

To ensure the effectiveness of the faulty science questions in challenging top-performing LLMs, I tested the dataset against leading APIs, including **Gemini 1.5 Flash**, **ChatGPT-4**, and **GPT-4o**. These models represent state-of-the-art capabilities in natural language understanding and reasoning, making them ideal candidates for evaluating the dataset's robustness. Preliminary testing revealed significant variation in the models' performance, with many questions successfully exposing reasoning flaws or limitations.

Dataset Composition

The final dataset comprises 200 carefully curated faulty science questions, categorized as follows: **Many subtopics in these topics.**

- **Physics**: Problems introducing false assumptions about mechanics, quantum phenomena, or thermodynamics.
- **Materials Science**: Misleading scenarios about biomaterials, ceramics, and material testing methods.

- **Chemistry**: Faulty problems involving incorrect chemical reactions, stoichiometry, or molecular structures.
- **Energy & Fuels**: Questions with errors in energy efficiency, sustainability, or fuel properties.

Each question in the dataset is accompanied by:

- The original valid question and its correct solution.
- The faulty version of the question.
- An explanation of why the question is faulty.
- The classification label assigned during the self-verification process.

Experiments

I have conducted experiments on whole 5000 dataset and curated 200 final dataset.

Experiment 1: Evaluating the Impact of Temperature Settings on Faulty Question Identification

Objective

This experiment aimed to assess how temperature settings in LLM configurations affect the ability of Gemini API 1.5 Flash to detect faulty science questions. I focused on three temperature levels: 0.2 (low, deterministic), 1.0 (balanced), and 1.2 (high, highly random), to analyze their impact on accuracy, consistency, and response clarity.

Methodology

A curated subset of 200 faulty science questions was selected, covering topics from physics, chemistry, materials science, and energy. Each question was tested across the three temperature settings. For each test, Gemini was tasked to:

1. Answer the question.
2. Identify whether the question was faulty, solvable, or ambiguous.
3. Provide a reasoning-based explanation.

As the API is free and I have 15 request per minute limit.

Each question was tested three times per temperature setting, and results were evaluated based on:

Accuracy: Percentage of correctly identified faulty questions.

Fault Detection Rate: Proportion of faulty questions flagged correctly as faulty.

Findings

1. **Fault Detection Rate**:

    - **0.2**: Fault detection rate was **55%**, showing the model’s reliability at low randomness.
    - **1.0**: Dropped to **42%**, as variability in reasoning led to occasional misinterpretation of faulty questions as solvable.
    - **1.2**: Fault detection rate further declined to **40%**, with hallucinated content affecting the model's judgment.

At 1.2 there were more hallucinations ig.

Discussion

The experiment revealed a clear trade-off between randomness and performance in faulty question detection. At **0.2**, the model delivered precise, consistent, and accurate results, making it ideal for tasks requiring deterministic reasoning. At **1.0**, responses exhibited greater diversity but at the cost of reduced accuracy and consistency. At **1.2**, randomness significantly impacted reliability, resulting in hallucinations and reduced fault detection rates.

The findings indicate that **0.2** is the optimal temperature for evaluating faulty science questions, as it ensures high accuracy and logical coherence. **1.0** may be suitable for tasks requiring some level of creativity, but **1.2** introduces excessive variability, rendering it unsuitable for logical reasoning tasks.

Here is the excel sheet link for 3 temperatures : temperature

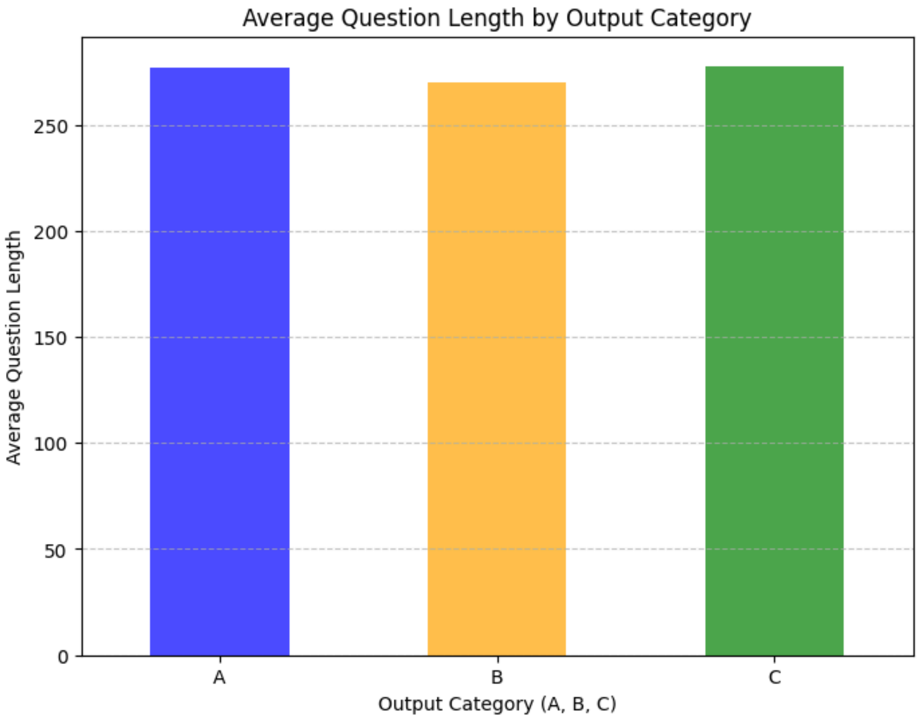
Experiment 2: Impact of Question Length on Fault Classification

Objective

This experiment aims to explore how the length of a question influences its classification as A, B, or C during the self-verification process by Qwen 32B. Specifically:

- **A**: Faulty but easy to identify as such.
- **B**: Genuinely faulty and challenging to identify.
- **C**: Valid and solvable, despite being labeled as faulty.

I investigate whether question complexity (as approximated by length) correlates with these classifications, particularly focusing on how often longer or shorter questions are classified as B, the most challenging category.



The graph shows that questions classified as B (genuinely faulty and challenging) have a slightly smaller average length compared to A (faulty but easy) and C (valid and solvable). **This result may indicate that question complexity, as measured by length, does not always directly correlate with the challenge level for LLMs.**

Sheet Link: **5000**

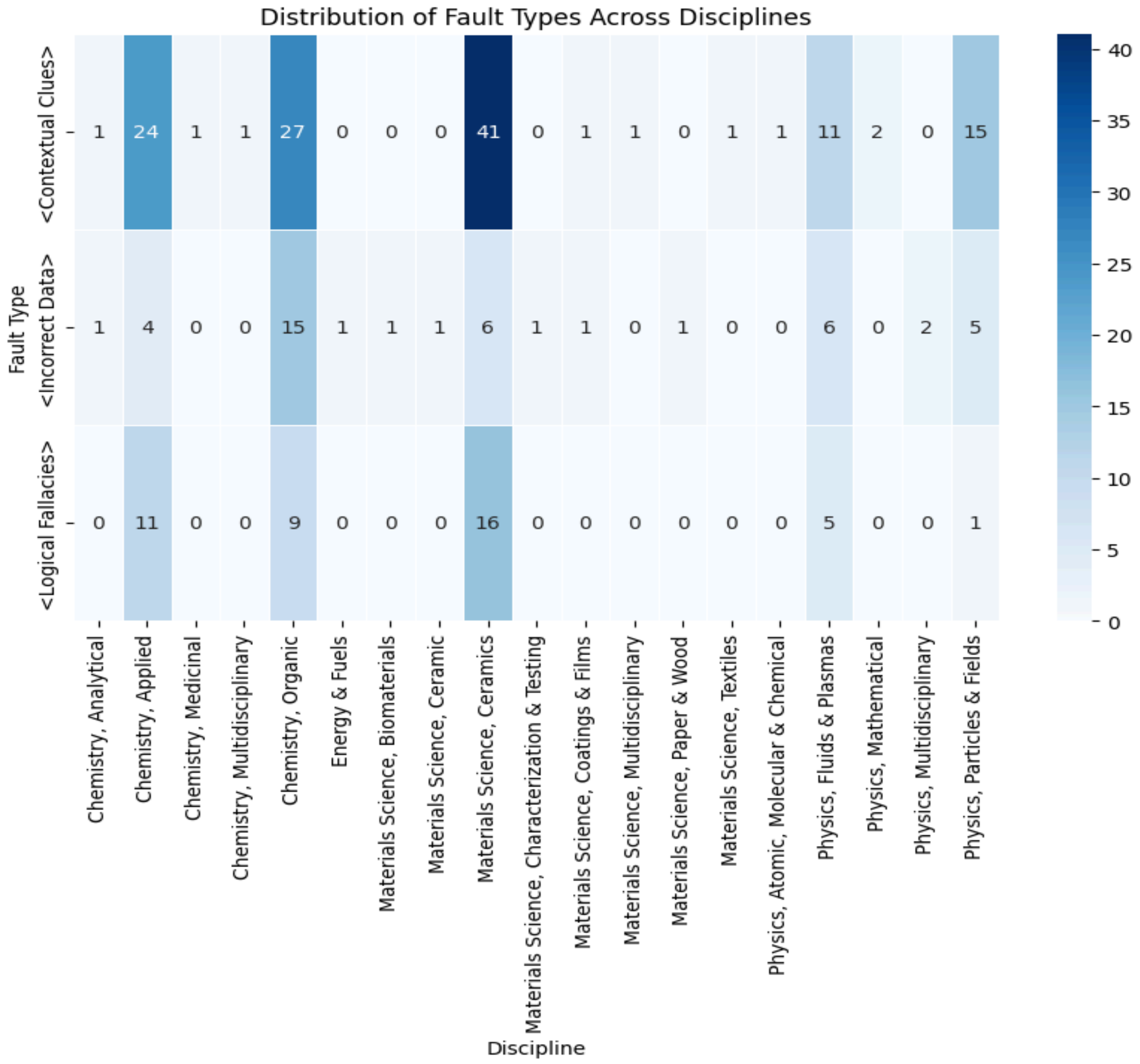
Experiment 3: Distribution of Fault Types Across Disciplines

Objective

The goal of this experiment was to analyze the distribution of different fault types across various scientific disciplines in our faulty science dataset. By visualizing the occurrence of faults categorized as `Incorrect Data`, `Logical Fallacies`, and `Contextual Clues` within each discipline, I aimed to identify patterns and trends that might affect reasoning performance. I classified it by Gemini 1.5 flash.

Dataset Preparation:

- The dataset consisted of faulty science questions, with each question categorized under a `Discipline` (e.g., Chemistry, Physics, Materials Science).
- Fault types were assigned to each question using the Gemini API 1.5 Flash model. The fault types included:
  - `Incorrect Data`: Factual errors or contradictions.
  - `Logical Fallacies`: Flawed reasoning or invalid inferences.
  - `Contextual Clues`: Missing or misleading contextual information.



Analysis and Insights

1. **Contextual Clues Dominance**: The most frequent fault type across disciplines, especially in `Materials Science, Ceramic` (41) and `Chemistry, Multidisciplinary` (27). This indicates that many questions obscure or manipulate context, challenging comprehension.
2. **Domain-Specific Fault Trends**: `Materials Science, Ceramic` and `Chemistry, Multidisciplinary` show the highest fault counts, reflecting their complexity and diverse reasoning challenges.
3. **Logical Fallacies**: Significant in `Materials Science, Ceramic` (16) and `Chemistry, Multidisciplinary` (11), indicating reasoning flaws are prominent in technical and interdisciplinary domains.
4. **Incorrect Data**: Common in `Chemistry, Multidisciplinary` (15) and `Materials Science, Biomaterials` (6), highlighting the importance of factual accuracy in these areas.

Key Takeaways

- `Contextual Clues` dominate across disciplines, making them a critical challenge for LLM reasoning.
- `Materials Science` and `Chemistry` exhibit the most diverse and frequent fault distributions, reflecting their inherent complexity.

Sheet Link : fault type

Experiment 4: LLM Accuracy Across Fault Types

Objective

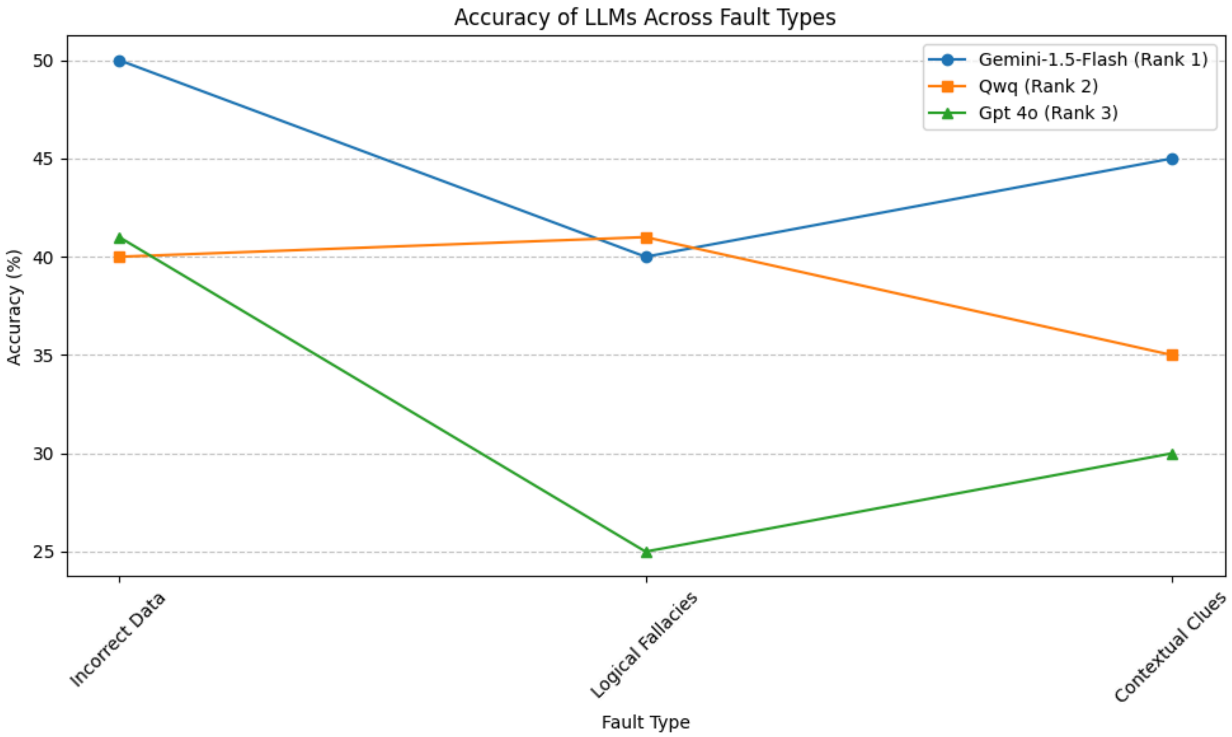
The objective of this experiment was to compare the accuracy of three top-performing LLMs Gemini-1.5, Qwq, and GPT-4 on identifying and resolving faulty science questions classified into three fault types: `Incorrect Data`, `Logical Fallacies`, and `Contextual Clues`.

Dataset:

The faulty science dataset was used, containing questions classified into three fault types:

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

LLM's Evaluated are :- Gemini-1.5-Flash, Qwq, GPT-4o



Analysis

1. **Gemini-1.5-Flash**:
    - Achieved the highest accuracy across all fault types, with strong performance on `Incorrect Data` (40%) and `Contextual Clues` (45%).
2. **Qwq**:
    - Performed well for `Incorrect Data` (45%) but struggled with `Contextual Clues` (35%).
3. **GPT-4**:
    - Showed declining accuracy for `Logical Fallacies` (30%) but improved slightly for `Contextual Clues` (35%).
4. **Fault Type Trends**:
    - `Logical Fallacies` were the most challenging for all models, while `Incorrect Data` had the highest accuracy overall.

This analysis suggests focusing future improvements on enhancing LLMs' ability to resolve logical inconsistencies and ambiguous contexts.

Experiment 5: Impact of Providing Hints on Fault Detection Accuracy

Objective

To evaluate how providing hints influences the performance of **Qwq** and **Gemini-1.5 Flash** 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**:

    - **Qwq**
    - **Gemini-1.5 Flash**

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**:

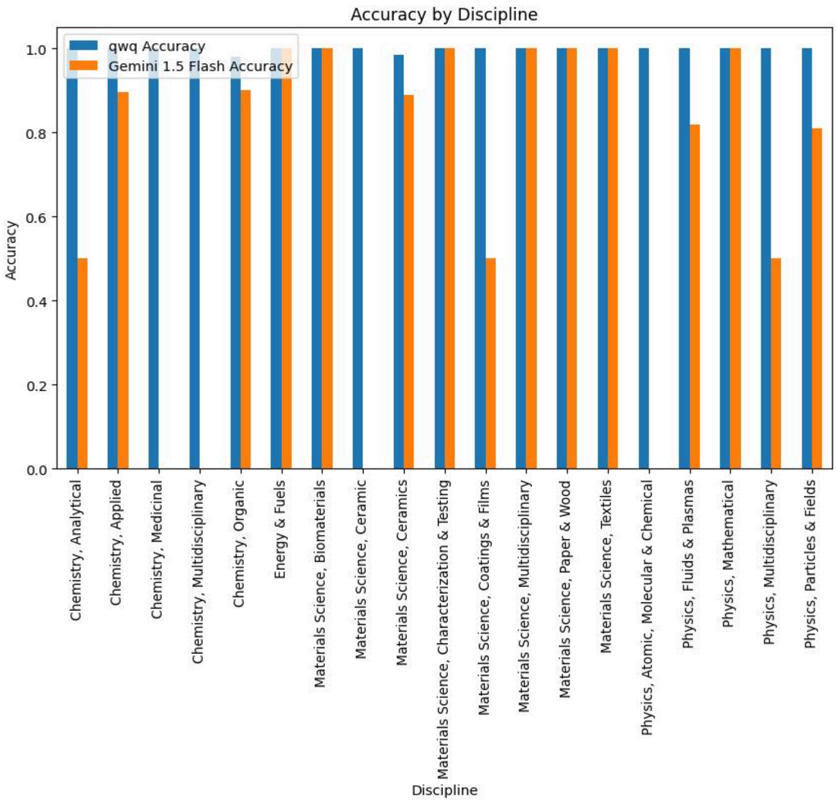
    - **Qwq** consistently outperformed **Gemini-1.5 Flash** across most disciplines.
    - Notable performance differences were observed in `Chemistry` and `Materials Science`, where Qwq demonstrated higher accuracy.

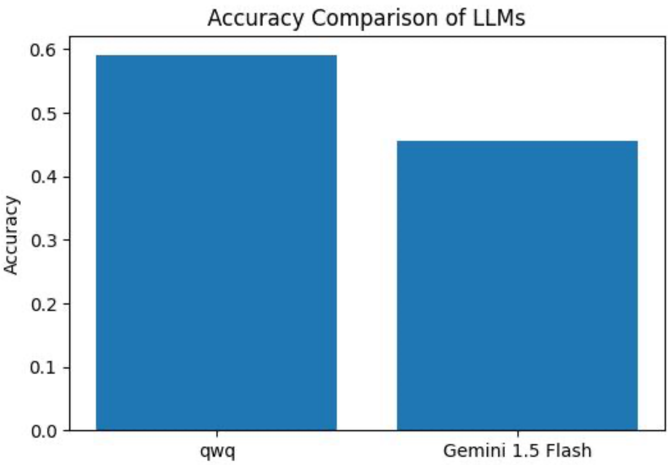
2. **Overall Accuracy**:

    - Qwq achieved an average accuracy of **~60%**, while Gemini-1.5 Flash achieved **~50%**.
    - The performance gap indicates Qwq's stronger ability to leverage hints for fault detection.

3. **Discipline-Specific Observations**:

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





Conclusion

This experiment demonstrates that:

- **Qwq** benefits more from hints, achieving higher accuracy than **Gemini-1.5 Flash** 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.

Future work could explore more structured hints or task-specific fine-tuning to enhance accuracy further.