Name: Sai Chandana Avvaru
PSU ID: 972317206
PSU Email: sfa5870@psu.edu

CSE 584: Final Project Report

Introduction –

The questions which are flawed, though perhaps not immediately appearing at fault, but containing underlying flaws that render them intractable or absurd, are actually the most useful in ascertaining the strength and reasoning powers of LLMs because they highlight their failings with regard to logical thinking, domain-specific knowledge, and even the ability to identify an ill-posed problem. These gaps need to be overcome if one wants LLMs to be dependable in real-world applications where precise comprehension and reasoning are required.

I created a database of incorrect questions with the help of Hugging Face's SciQ dataset, which consists of over 10,000 scientific questions. I requested GPT-3.0 to choose 2,000 lengthy, intricate, and logically sound questions from this database using prompt engineering. GPT-4.0 was then tasked with adding faults to these questions so that other LLMs that I would use to experiment with could not answer them. The faulty questions were then applied to different LLMs in batches of 100. I manually found questions that had been answered wrongly without fault detection. Only the faulty questions that LLMs answer poorly were included in the final database.

This report explains the creation of the problematic question dataset, tests conducted to evaluate LLM performance, and lessons gained from their mistakes. The results not only point out areas where LLM reasoning must be strengthened, but they also provide a valuable resource for future studies on boosting model resilience.

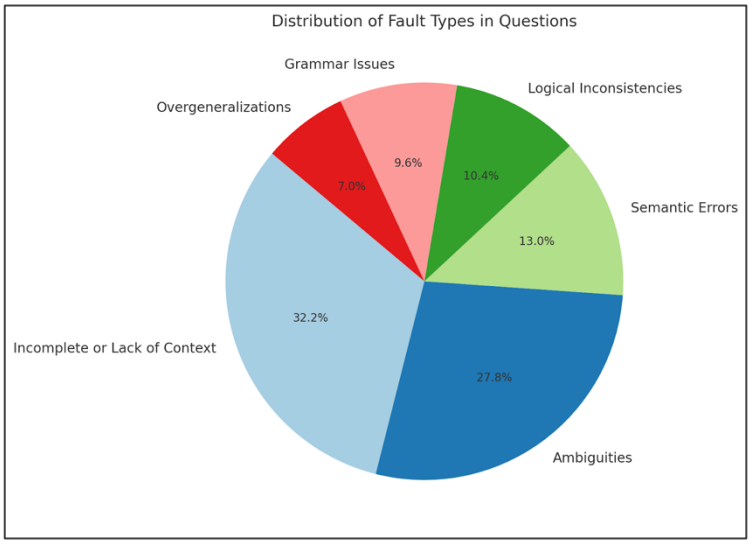
RESEARCH QUESTIONS –

1. What types of faults are most likely to go undetected by current top-performing LLMs?

After extended experimentation with batches of faulty questions tested against various LLMs, I noticed three major fault types that passed through the LLMs undetected - **logical inconsistencies, semantic errors, and ambiguities**. These faults attack the deficiencies in logical reasoning, pattern recognition, and contextual understanding of the LLMs to get confident yet flawed responses.

I divided the faulty questions into batches, then went through each of the LLM responses manually to find that:

- Logical Inconsistencies accounted for many confidently incorrect answers, especially in disciplines like physics and biology.
- Semantic Errors were particularly effective at misguiding LLMs when scientific terminology was subtly abused.
- Ambiguities highlight how LLMs do more over-interpreting when it comes to ambiguous questions, rather than looking for clarification.



I also found a few patterns of failure of the LLMs –

- Questions designed to exploit surface-level reasoning and pattern matching were particularly effective at exposing LLM weaknesses.
- Faults that combined multiple categories, such as ambiguous questions with semantic errors, were more likely to yield incorrect responses.

2. What are the common patterns in incorrect LLM responses to faulty questions?

While I was manually finding questions that had been answered wrongly by the LLMs without fault detection, I observed some key patterns of bad questions often generate bad LLM responses such as: overconfidence, surface reasoning, and inability to detect contradiction. These reflect deep limitations within current LLM architectures to focus on generating plausible responses rather than deeply comprehending or critiquing the input.

1. **Overconfidence:** LLMs frequently provide detailed, confident responses even when the question itself is flawed. This overconfidence stems from their design to maximize fluency and informativeness, often prioritizing coherence over accuracy. LLMs generate answers assuming that the input is correct. They do not question whether the premise is valid or possible, leading to authoritative but incorrect responses.
    - Example: "Give a precise time frame, in milliseconds, for the full depolarization of a cellular membrane."
    - LLM Response: "Depolarization lasts 1-2 milliseconds," firmly establishing a time frame without taking a moment to consider that such precision cannot be universally established for all cell types.

    In this case, the LLM took the question to be valid and used its pattern memory from training data to arrive at an answer. The over-confidence thus misleads the user in particular in critical reasoning-type questions or for scientifically accurate information. A human expert would argue with the premise, not attempt to answer it.

2. **Shallow Reasoning:** Many times, LLMs end up matching the pattern of the input rather than its logical analysis. This behavior arises due to the model's reliance on statistical

correlations learned during training, rather than true comprehension regarding the meaning of the question. LLMs identify familiar structures in the question and attempt to answer based on linguistic patterns rather than verifying the logical validity of the input.

- Example: "Birds convert food into molecules small enough to be absorbed by epithelial cells of what?"
- LLM Response: "The intestines," focusing on the mention of "epithelial cells" and "food" while ignoring the biologically flawed phrasing of the question.

The response is a possible relationship, but it does not consider the illogical framing of the question, which confuses processes. Surface-level reasoning limits LLMs in handling nuanced or complex questions, especially in scientific or technical fields where deep logical analysis is required.

3. **Failure to Detect Contradictions:** LLMs do not have any mechanism to cross-check the premises of a question against their internal knowledge. Hence, they cannot detect a contradiction, or an impossible event stated in the input. The model accepts any contradictory or scientifically incorrect statement as a fact and then prepares an answer based on the wrong premise.

   - Example: "Reptiles are endothermic because their heat comes from internal sources."
   - LLM Response: "Endothermic animals regulate heat internally," making a confident agreement to a premise without any notation that reptiles, therefore, must be ectothermic, meaning that they get their heat from environmental sources.

   Such failure occurs because here the model does not reason out the scientific definition for "endothermic" and "reptile" but focuses on pattern-matching like "endothermic animals" with "regulate heat internally." This is an especially dangerous problem in the sciences and engineering, where the underlying logic may be pivotal. LLMs can't discern between logical and non-logical premises, unless told otherwise.

Hidden Patterns Explaining These –

- *Prejudice in Training Data:* LLMs are trained on massive data, mostly of correct facts. They therefore become biased to always believe in the validity of input and never question it.
- *Lack of Logical Validation:* LLMs are not inherently equipped with a mechanism of reasoning systems for logical consistency or challenging of assumptions.
- *Objective Design:* The models are optimized for the generation of fluent and plausible responses but not to critically evaluate the input given. Their training favors "what sounds right" over "what is logically or scientifically correct."

Implications –

- These patterns highlight the need for improvements in LLM architectures, such as integrating reasoning modules or logical consistency checks.
- Critical applications require framing inputs with utmost care and evaluating outputs separately since model confidence does not necessarily translate to accuracy.

3. Can prompt engineering reduce the likelihood of LLMs being fooled by faulty questions?

Prompt engineering, in cases of faulty questions helps reduce failures because it makes the model find errors and inconsistencies before coming up with an answer. Although it cannot absolutely surmount inherent limitations in reasoning, it seriously improves the chances that the model will question faulty premises. But this can be primed through an instruction that the LLM will look for faults or inconsistencies first before attempting to provide an answer. In this regard, the model will analyze the question, prioritize error detection, and respond accordingly, in context.

Consider the original prompting of the LLM with the faulty question, just plainly posing a faulty question for it to answer:

- Original Prompt: **"What happens to a rock when under stress without shape change?"**
- LLM Response: **"It undergoes elastic deformation."**

The response assumes the premise is valid and provides a direct answer based on patterns in its training data. Now, when I give another specific prompt immediately to the LLM -

- Engineered Prompt: **"Identify any logical errors in the question and then answer."**
- LLM Response: **"The question assumes all rocks exhibit the same behavior under stress, which is incorrect. Rocks behave differently depending on their composition and external conditions."**

By reframing the prompt, the model evaluates the question's assumptions, identifies flaws, and provides a more nuanced response. I also got to know that by asking the model to "analyze for errors," prompt engineering shifts the model's task from simply answering to critiquing. This is my prompt -

- Original Prompt: **"Critique this question: 'Reptiles are endothermic because their heat comes from internal sources.' Then answer it."**
- LLM Response: **"This question contains an error: reptiles are ectothermic, not endothermic. They rely on external sources for heat."**

Also, prompting the model to clarify ambiguities before answering improved its ability to deal with vague or incomplete questions. It was done this way -

- Prompt: **"Identify ambiguities in this question before answering: 'What is the most environmentally friendly chemical?'"**
- LLM Response: **"The question is ambiguous, as it doesn't define what 'environmentally friendly' means or the context of usage. Do you mean low toxicity, biodegradability, or carbon footprint?"**

In another experiment, I observed that structured prompts encourage logical consistency evaluation in models -

- Prompt: **"Assess this question for logical consistency: 'What fertilizer equals the distance and effort of photosynthesis?'"**
- Response: **“Question is illogical because it combines physical efforts and chemical processes. 'Effort' does not relate to the fertilizers in photosynthesis.”**

Even with explicit prompts, I observed that the LLMs sometimes fail to recognize fundamentally flawed premises when they are framed in complex ways.

- Prompt: **"What changes increase torque for a perpetual motion?"**
- Response: **"The concept of perpetual motion is not scientifically possible. However, torque can be improved through the use of gyroscopes."**

Even when prompted to critique, the model responds with which is again incomplete reasoning. The effectiveness of prompt engineering also greatly relies on how specific and clear the instructions are. Poorly designed prompts do not achieve the required result. LLMs will not always interpret prompts as intended, especially in edge cases when the faulty premise is subtle or domain specific.

4. Are there limitations in LLMs' logical reasoning that contribute to their failure on faulty questions?

While I was manually finding questions that had been answered wrongly by the LLMs without fault detection, I got to observe the LLM's responses. I saw that LLMs do suffer from important shortcomings regarding logical reasoning; these are some of the reasons behind their failure with flawed questions:

1. Inability to Cross-Verify Premises: LLMs do not dynamically compare the premises of a question with known facts.
    - Example: **"All chemical changes involve a transfer of electrons, even in nuclear reactions."**
    - LLM Response: **"The carbon allotrope diamond has the highest melting point," out of context-specific conditions.**
2. Reliance on Patterns: LLMs rely on linguistic patterns rather than in-depth logical analysis, which results in plausible but wrong answers.
    - Example: **"Give a precise timeframe, in milliseconds, for the depolarization of a membrane, without specifying the type of membrane or cell."**
    - LLM Response**: "Membrane depolarization typically lasts 1-2 milliseconds, particularly in neurons or excitable cells, though this can vary depending on the cell type and mechanism.**
3. Lack of Ability to Recognize Logical Fallacies: LLMs reinforce flawed premises with authoritative responses.
    - Example: **"Oxygen forms compounds only with fluorine."**
    - LLM Response: **"Oxygen reacts with fluorine to form OF_2," instead of critiquing the false premise.**
4. Context Dependence: Answers change profoundly with the framing of the question, revealing illogical thought processes.

Explicit directions such as "Analyze this question for logical fallacies" can help the performance by specifically guiding LLMs to critique the input. Without explicit guidance, these models do not have dynamic reasoning or cross-verification of information, leading to many of their errors in logic.

5. What roles do domain-specific knowledge and logical consistency play in LLMs' susceptibility to faulty questions?

While I was manually finding questions that had been answered wrongly by the LLMs without fault detection, I observed that large language models are very sensitive to both the breadth and depth of domain-specific data in their training sets. Their capability of giving correct answers to questions depends on how a topic is represented and contextualized in their training. In less well-represented areas, such as nuclear physics or interdisciplinary topics, LLMs are most likely to give incorrect responses or overconfident incorrect responses.

Example:

- Faulty Question: " **All chemical changes involve a transfer of electrons, even in nuclear reactions, correct?"**
- LLM Response: **" Overgeneralization. Nuclear reactions involve changes in atomic nuclei.**

The scarcity of training data for some domains makes the possibility of LLMs misinterpreting or not recognizing faults in questions even higher. Logical coherence defines how well LLMs can assess the validity of a question's premise. LLMs rarely question flawed premises, even in well-understood domains.

Example:

- Faulty Question: **" Endothermic term given to animals like reptiles means that their heat comes from internal sources?"**
- LLM Response: **"Endothermic animals regulate heat internally," failing to recognize that reptiles are ectothermic.**

LLMs cannot dynamically cross-check premises against established logical frameworks.

Example:

- Faulty Question: **" All chemical changes involve a transfer of electrons, even in nuclear reactions, correct? "**
- LLM Response: **"Electron transfer is involved in chemical reactions," overlooking the contradiction with nuclear reactions.**

Models prioritize linguistic patterns over logical consistency, often yielding responses that are plausible yet incoherent.

Example:

- Faulty Question: **" What adjustments used in guidance systems to indicate directions in space must have an angular momentum that constitutes about ten percent of all what?"**
- LLM Response: **"Torque can be changed out with gyroscopes," not taking into account that there is no such thing as perpetual motion.**

When domain knowledge is sparse, the introduction of logical fallacies can go unnoticed more easily.

Example:

- Faulty Question: **" Because of electromagnetic waves of matter, the idea of well-defined orbits gives way to their speed always being constant; this constant is what?"**
- LLM: **" Probability - The question confuses two distinct concepts: quantum mechanics (probability clouds) and electromagnetic wave properties (speed of light).".**

6. Can faulty question datasets improve the evaluation metrics for benchmarking LLMs?

Indeed, faulty question datasets will substantially raise the bar for benchmarking LLMs on the tasks of detection and response to flawed premises, ambiguous phrasing, and logical inconsistencies. While traditional benchmarks would require either correctness or fluency, faulty question datasets allow insight into an LLM's capabilities of reasoning, robustness, and critiquing the input provided to them.

Faulty questions explicitly test logical consistency by presenting flawed premises.

- Example: **"All chemical reactions involve electron transfer, even in nuclear reactions."**
  - Traditional metrics: Measure how well the LLM discusses chemical and nuclear reactions.
  - Flawed dataset metric: Measure whether the LLM recognizes that the question contains an overgeneralization and therefore discredits the question's premise.

Faulty datasets examine whether an LLM is able to find and mention ambiguities, inconsistencies, or semantic mistakes before it makes a response.

- Example: **"Reptiles are endothermic because their heat comes from within themselves."**
  - A good LLM should flag the logical mistake (reptiles are ectothermic) and not provide an answer with a lot of confidence.

Vague questions test an LLM's ability to ask for clarification or provide responses conditionally.

- Example: **"What is the most environmentally friendly chemical?"**
  - A good LLM should detect vagueness (e.g., what is meant by "environmentally friendly") and ask for clarification.

The faulty datasets containing subtle flaws help the researchers determine whether LLMs are relying on remembered patterns or understanding the question.

Faulty datasets expand beyond accuracy to test reasoning, fault detection, and ambiguity handling, providing a more comprehensive assessment. These datasets reveal weaknesses in LLM training, such as susceptibility to adversarial inputs or logical inconsistencies, guiding improvements in model development. They assess not only if the model answers correctly, but how well it explains its reasoning or critiques the question. Models trained or tested with flawed questions become much better at error detection, increasing their usefulness in critical applications.

7. Are larger models more resilient to faulty questions than smaller models?

While I was manually finding questions that had been answered wrongly by the LLMs without fault detection, I observed that larger models are generally more resistant to defective questions than smaller models are. This robustness follows naturally from several factors in both their design and training; this, however, is relative, as larger models also are challenged when it comes to addressing subtle logical fallacies and ambiguities.

1. **Large Models, including ChatGPT-4o, Gemini, Qwen-72B:**

   Large models handle faulty questions well because they have been trained to infer a lot from the given texts and, therefore, are better suited for open-ended or ambiguous tasks.

   - *Robustness to Faulty Questions:* Large models can put up with ambiguity much better, and their responses to badly structured prompts sound more natural. For instance, ChatGPT-4o answered contextually appropriate questions that were either too specific or too vague, such as processing millisecond-scale biological processes using plausible timescales. Gemini and Qwen-72B also exhibited very good inferential performance, drawing logical conclusions even when there were gaps or ambiguities in the question.
   - *Generalization and Accuracy:* Large models tended to recognize when questions did not make sense or were deceptive. For example, responses such as "This question is nonsensical" or further redirecting to related topics demonstrate an ability to critically assess the input. They are more likely to give detailed explanations and elaborations, reflecting their more sophisticated training across varied datasets.
   - *Performance Across Disciplines:* These models worked exceptionally well in areas such as biology and environmental science, where it could contextualize poorly or broad questions to give meaningful answers.

2. **Small Models, including Meta-Llama, Nvidia-Llama:**

   Small models are not as robust against defects in the input structure and present responses that are less informative or contextually relevant in a query that is complex or multidisciplinary.

   - *Flaws with Faulty Inputs:* Models with a small size, such as Meta-Llama and Nvidia-Llama, had a weak performance in terms of the ambiguity of questions, mostly providing incomplete or overly simplistic answers. For instance, responses have tended to be iterative of the question itself or superficial when the input had not contained enough context.
   - *Limited Generalization:* They were rather weak in handling complex questions which required integration of information across multiple domains. Rather than resolving the ambiguities, irrelevant or partially correct answers had been provided.
   - *Discipline-Specific Weaknesses:* Regarding the smaller models, these underperformed on more nuanced reasoning disciplines, including biochemistry or physics. Inability to infer missing details or reframe faulty inputs ultimately limited their effectiveness.

Related Work –

LLM performance in subject-specific domains has been evaluated in large part using science datasets like SciQ and ARC (AI2 Reasoning Challenge). Specifically, SciQ offers a vast array of

well-constructed multiple-choice science questions along with explanations. These datasets do not include flawed questions, even when they assess factual knowledge. My project expands the usefulness of SciQ to incorporate robustness testing by utilizing it as a foundation for creating incorrect questions. By fusing aspects of subject-specific testing, logical reasoning evaluation, and adversarial question design, my study offers a fresh viewpoint on the limits of LLMs. Additionally, it provides a dataset for future studies aimed at enhancing defect detection and model resilience.

Dataset Design and Collection –

The dataset developed for this project is a subset of curated, faulty questions in science that are targeted to assess the reasoning abilities of state-of-the-art LLMs. In this section, the process used to design and collect this dataset is described in steps, along with its structure and key features.

Data Source: The foundation of this dataset is the SciQ dataset, a large repository of more than 10,000 science questions ranging over biology, chemistry, physics, and environmental science. SciQ was chosen because it is the broadest and deepest, with high-quality, logically sound questions that range over a wide swath of scientific topics. These questions served as the base for generating complex and logically correct questions.

Data Augmentation and Fault Induction:

1. Complex Question Selection:
    - Use GPT-3.0 for advanced prompt engineering to identify long, complex, and logically correct questions from the SciQ dataset.
    - From over 10,000+ questions, 2,000 questions were selected based on their capability to test fine-grained reasoning and understanding.
2. Fault Induction:
    - These questions were fed into GPT-4.0 and were instructed to subtly change these questions to introduce a logical or semantic fault. This would be done in a manner that the question sounded correct upon first reading but had some sort of error that made it unsolvable or misleading.
    - Fault examples:
        - Ambiguity in phrasing
        - Contradictory or nonsensical premises
        - Poor grammar leading to misinterpretation.

The faults were introduced deliberately, ensuring that the questions:

- Appeared Valid at First Glance: The modified questions were designed to resemble logically correct questions, making them harder for LLMs to detect as faulty.
- Contained Subtle Faults: Each question was altered in a way that introduced logical inconsistencies, semantic errors, or ambiguity.

Manual Validation: To ensure the quality of the faulty questions, each question was manually reviewed for:

- Clarity: Ensuring the question appears valid to a casual reader.
- Fault Relevance: Confirming the introduced fault significantly affects solvability or logical coherence.
- Discipline Consistency: Aligning the question's theme with its scientific discipline.

Dataset Structure: There are five attributes in the final dataset:

1. Discipline: Indicates the scientific field, such as physics, chemistry, or biology.
2. Question: This includes the corrected incorrect question.
3. Reason for Fault: Explains the error that was inserted into the query.
4. LLM Evaluated: Shows the best-performing LLM that was tested using the question (e.g., Claude-3-Opus, GPT-4.0).
5. LLM Response: Describes the LLM's reaction, emphasizing whether it overlooked the error.

Experimentation –

For the research, a variety of cutting-edge LLMs have been used to examine this dataset of flawed science queries:

1. **GPT Chat 4.0:** A well-known model created by OpenAI that is commended for its capacity for reasoning and discussion.
2. **Gemini:** An exceptionally competent LLM who has honed a wealth of resources for difficult assignments.
3. **Claude.ai:** Anthropic created this platform, which focuses on morally sound and reliable AI solutions.
4. **Meta-Llama:** Meta's cutting-edge model, tailored for both general and domain-specific activities.
5. **Qwen-72B:** A sizable LLM renowned for its scalability and logic.
6. **Nvidia-Llama:** An algorithm that makes use of NVIDIA's sophisticated computational optimizations.
7. **Coral:** A more recent LLM that underwent generalization and robustness testing.
8. **Perplexity:** Limited testing, some fault detection but lacked breadth.

Experimental Setup:

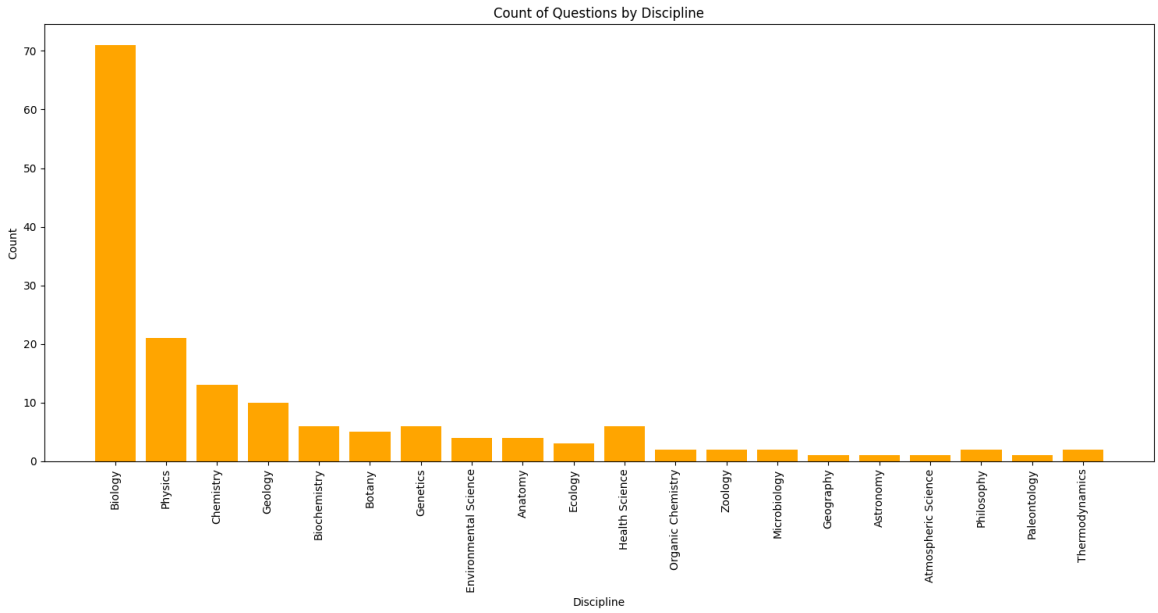
1. Question Batches: To ensure systematic testing of the dataset across several models, the problematic questions were separated into batches of 100. To assess their performance, the LLMs were shown each batch in turn.
2. Testing Protocol: The identical flawed questions were asked to every LLM. No further context or cues were given to the models to help them recognize the errors; thus, their replies were solely dependent on their ability to reason. The responses were noted for further review.

3. Evaluation Metrics:
    - Fault Detection: If the query was recognized as flawed by the model.
    - Answer Validity: Whether, despite the question's flaws, the model produced a logical or scientifically sound response.
    - Error Types: The kinds of mistakes that were made in the answers, such as logical inconsistencies, factual errors, or failure to identify faults.
4. Manual Verification: Following receipt of the replies, each was personally examined to: Confirm if the LLM detected the error. Determine the sort of success or failure in answering the incorrect question. Responses that were deemed unsuccessful were those who disregarded the error or boldly gave a wrong response.

Dataset Composition –

The dataset consists of 127 distinct faulty questions that have been meticulously constructed and verified to test LLMs' capacity for reasoning.

1. **Diversity of Disciplines**:



- Numerous scientific fields, including biology, chemistry, physics, and environmental science, are represented among the flawed questions. Because of this diversity, the dataset is guaranteed to assess LLMs' thinking over a broad spectrum of knowledge fields.
- Examples:
    - **Biology**: "The cell cycle's primary regulatory mechanism is attributed solely to DNA polymerase activity."
    - **Physics**: "What adjustments used in guidance systems to improve torque lead to perpetual motion?"'

2. **Types of Faults Introduced:**

   Each question has an embedded error that makes it impossible or misleading to answer, yet it has a superficial appearance of being valid. The faults can be categorized as:

   i. **Logical Inconsistencies**:

   - **Question**: *"What type of energy of a system is the sum of the kinetic or maybe potential energies of its atoms or maybe molecules?"*
     - **Reason**: Confuses the definition of internal energy with vague phrasing.
   - **Question**: *"When a metal is oxidized or maybe a nonmetal is reduced in a redox reaction, what is the resulting compound called?"*
     - **Reason**: Misrepresents redox reactions with misleading phrasing.

   ii. **Ambiguities**:

   - **Question**: *"What type of molecules are secreted by the cell in local signaling?"*
     - **Reason**: Lacks specificity, leaving multiple interpretations.
   - **Question**: *"What type of organisms has many different specialized cells that work together to carry out life processes?"*
     - **Reason**: Ambiguous phrasing makes it unclear.

   iii. **Semantic Errors**:

   - **Question**: *"Why are females influenced by testosterone during development?"*
     - **Reason**: Misrepresents the role of testosterone in female development.
   - **Question**: *"Yeasts are single-celled fungi. About 1,000 species are recognized, but the most common species is Saccharomyces cerevisiae, which is used in this?"*
     - **Reason**: Poor phrasing and misuse of scientific context.

   iv. **Grammar Issues**:

   - **Question**: *"What thin layer of air acts as a barrier to prevent cool air from mixing with warm air in the stratosphere?"*
     - **Reason**: Poor sentence construction creates confusion.
   - **Question**: *"What type of matter in the spinal cord has the appearance of an ink-blot test?"*
     - **Reason**: Sentence structure affects clarity.

   v. **Incomplete or Lack of Context**:

   - **Question**: *"What tissue class includes epithelial, nervous, muscular, and connective?"*
     - **Reason**: Lacks additional detail for better understanding.

- **Question**: *"What type of radiation from the sun reaches Earth across space, striking everything on Earth's surface?"*
    - **Reason**: Missing clarity on the type of radiation expected.

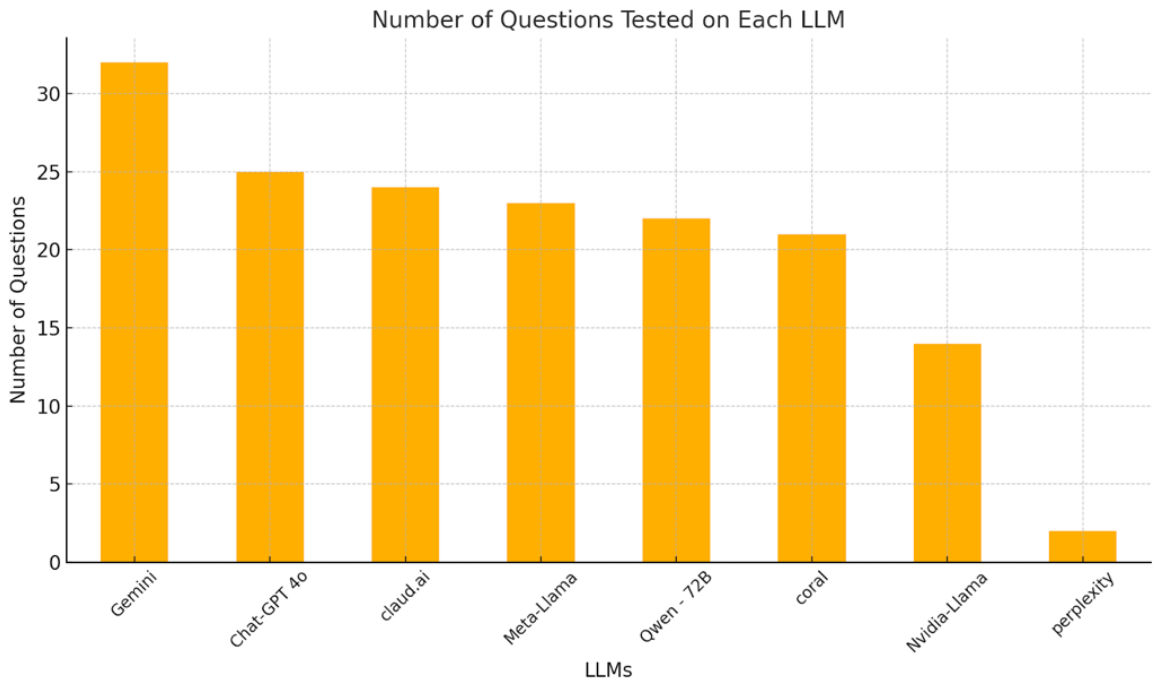
vi. **Overgeneralizations**:

- **Question**: *"What type of substance has a fixed chemical composition and characteristic properties?"*
    - **Reason**: Oversimplifies the definition of pure substances.
- **Question**: *"What term refers to the friction of fluid, within itself and its surroundings?"*
    - **Reason**: Generalizes fluid dynamics with insufficient specificity.

3. **Repetition Avoidance**:

- A conscious effort was made to avoid having duplicate or similarly worded questions within the dataset.
- The 127 unique questions represent independent test cases, each of which targets different weaknesses in LLM reasoning.

<u>Results Summary –</u>



- <u>Chat-GPT 4.0:</u> Erroneous on 25 faulty questions; confident but missed subtle logical inconsistencies.
- <u>Gemini:</u> Erroneous on 32 faulty questions; struggled with semantic ambiguities, gave plausible but wrong answers.
- <u>Claude.ai:</u> Erroneous on 24 faulty questions; good with grammar faults, less effective on logic errors.

- Meta-Llama: Erroneous on 23 faulty questions; strong on straightforward faults, weaker on complex logic.
- Qwen-72B: Wrong on 22 flawed questions; detailed, but question flaws were missed.
- Nvidia-Llama: Wrong on 14 flawed questions; concise and with weaker fault recognition.
- Coral: Wrong on 21 flawed questions; strong fact-based reasoning, struggled with badly formulated problems.
- Perplexity: 2 flawed questions wrong; little testing involved, some fault detection without depth. Limited testing showed some capacity for fault detection but lacked breadth for robust evaluation.

Insights from Testing:

- Inconsistent Fault Recognition: Without identifying logical or semantic mistakes, most models confidently responded to incorrect queries. In contrast to overt grammatical errors, subtle errors were more successful in deceiving LLMs.
- Domain-Specific Weaknesses: Chemistry and biology fared better, but physics and multidisciplinary environmental science problems were particularly challenging.
- Room for Improvement: The findings emphasize the necessity of enhancing LLMs' reasoning processes and fault-detection skills.

Role of Unique Questions in the Dataset:

- Comprehensive Evaluation: This ensures that the dataset includes a comprehensive evaluation of LLMs across various reasoning tasks by having 127 unique faulty questions. It eschews redundancy, ensuring each test contributes new insights into what these models can and cannot do.
- Focus on Difficult-to-Detect Faults: Unique questions prevent LLMs from relying on pattern recognition or memorized responses. Each question tests a specific logical or conceptual flaw that demands real reasoning from the model.

Examples and Their Impact: Here are some examples to help grasp the breadth and intricacy of these questions:

1. **Question**: "Give a precise timeframe, in milliseconds, for the complete depolarization of a cellular membrane."
    - **Fault**: The question assumes an unreasonable level of precision, which is not applicable to biological processes due to variability.
    - **Impact**: Highlights LLMs' inability to question the premise or point out the unreasonable assumption.
2. **Question**: "Name the single, most environmentally friendly chemical compound."
    - **Fault**: There is no such universally accepted "most environmentally friendly" compound, making the question unanswerable.
    - **Impact**: Tests the LLM's understanding of open-ended or ill-defined scientific concepts.

Detailed Analysis of LLMs' Performance on Faulty Questions: The assessment of Large Language Models on the dataset of flawed questions revealed valuable insights about their reasoning, fault-detection capabilities, and robustness in general. A detailed breakdown of each of these dimensions of their performance follows:

1. **Fault Recognition -**
   - **General Performance:**
     - Most LLMs did not identify faults in a large percentage of the question, thus proving themselves vulnerable to subtle errors of logic, semantics, and grammar.
     - Questions with embedded semantic or logical fallacies proved particularly difficult, since LLMs often tried to provide direct answers rather than indicating the flaw.
   - **LLMs Specific Observations:**
     - *Chat-GPT 4.0:*
       - Given its assurance in providing responses, logical fallacies were often ignored.
       - Example: There is a question: “What fertilizer is equivalent to distance, effort, and time of photosynthesis?”, this question was answered as “Eutrophication” without acknowledging the preposterous lead.
     - *Claude.ai:*
       - Did better in finding grammatical faults but failed to handle deeper logical inconsistencies.
     - *Gemini:*
       - Did relatively well with simpler questions; however, as the questions get more complicated, its rate of fault detection goes down.
     - *Meta-Llama:*
       - Found out simple faults but gave logically inconsistent questions reasonable, but incorrect, answers.

2. **Response Validity -**
   - **Accuracy vs. Fault Handling:**
     - Even while delivering scientifically correct information, most LLMs lacked addressing the faulty nature of such a question.
     - It shows evidence of pattern-matching and shallow, surface-like reasoning rather than deeper logical understanding.
   - **Common Mistakes:**
     - Logical Contradictions: For the flawed physics question "What adjustments used in guidance systems improve torque for perpetual motion?", most LLMs responded confidently with either "gyroscopes" or "momentum", missing the point that perpetual motion is impossible.

- Semantic Misinterpretation: The question was "Name the single, most environmentally friendly chemical compound," for which the LLMs gave answers such as "Water (H_2O)," even though the question can't be answered because it's based on an unanswerable premise.

3. **Domain-Specific Performance -**

   - **Biology and Chemistry:**
     - LLMs performed relatively better in these disciplines, probably because of well-structured training data in these domains.
     - Mistakes of exact timing-for example, "Give a precise timescale, in milliseconds, for the full depolarization of a cellular membrane"-were overlooked many times.
   - **Physics:**
     - Questions with logical flaws or impossible situations were the most effective in misleading LLMs.
     - Example: Questions involving impossible gear ratios or mechanisms of perpetual motion always fooled models.
   - **Environmental Science:**
     - Interdisciplinary questions with ambiguous premises or poorly defined terms showed models' weaknesses regarding how well they could process open-ended, nuanced problems.

Insights from Manual Checking:

1. **Overconfidence in Faulty Answers:** Confidence and detail when answering flawed questions were well manifested in LLMs, seemingly oblivious to inherent flaws that might mislead a user seeking truthful information from LLMs.
2. **Susceptibility to Subtle Faults:** Subtle faults, such as semantic errors or slight logical inconsistencies, can easily fool an LLM.

   Example: The flawed biology question, "The cell cycle's major regulatory mechanism is due solely to DNA polymerase activity," was responded to without any mention of the incorrect attribution made in the question.
3. **Deficiencies in Reasoning:** The majority of LLMs also failed to cross-check their responses against the logical structure of the question. This is a serious deficiency in their reasoning powers regarding ill-defined or paradoxical situations.
4. **Lack of Discipline-Specific Accommodation:** Performance was strongly divided among subjects, but for physics and environmental science more so than for biology and chemistry.
5. **Over-Confidence in Patterns:** LLMS was found many a time sticking to the linguistic or structural pattern in developing an answer when required, thereby leading to their failures with non-standard/faulty question formats.

Key Findings and Implications:

1. **Common Failure Points:** Questions that combined interdisciplinary knowledge or required multi-step logical reasoning were particularly challenging. Semantic ambiguities and illogical premises were effective at misleading LLMs, revealing a need for improved fault-detection mechanisms.
2. **Importance of Manual Validation:** Manual checking played a critical role in verifying faults and identifying response patterns. Automated methods alone were insufficient for nuanced evaluation.
3. **Need for Enhanced Fault Detection**: Current LLMs lack robust mechanisms for the identification and handling of ill-posed or faulty questions. Equipping them with such would enhance their reliability in real-world applications.
4. **Dataset Insights:** The dataset of faulty questions collected on its own provides a good resource to test and enhance the robustness of LLMs. It identifies specific points where models fail, therefore acting like a roadmap for future development.

Conclusion –

This work points out some key gaps in the reasoning and fault-detection capabilities of state-of-the-art LLMs. The construction of a dataset containing faulty questions and the subsequent assessment of the performance of LLMs demonstrated quite clearly that even the most sophisticated of these models lack the ability to handle subtle logical inconsistencies, semantic errors, and ambiguities. These findings stress the need for developing better reasoning capabilities and strong fault-detection mechanisms to enhance LLM reliability. The curated dataset is useful because it forms a good benchmark, enabling future improvements and therefore making the path ahead much smoother to develop resilient and trusted AI systems.