Teja Sree Parasa
CSE 584: Machine Learning

December 6th, 2024
Final Project

Assessing Logical Reasoning in Large Language Models: Evaluating Performance on Faulty Scientific Questions

Introduction

Large language models (LLMs) like ChatGPT-4, Gemini-1.5-Pro, and Claude-3-Opus have become increasingly important in fields like education, healthcare, and scientific research. They're great at answering questions, solving problems, and reasoning through complicated topics. But they're not perfect. One major issue is that these models sometimes fail to catch subtle logical errors in questions, which can lead to answers that sound convincing but are actually wrong. Figuring out why this happens and how to fix it is really important for making AI more reliable.

For this project, I focused on testing how well LLMs can spot and reason through flawed questions. I created a dataset of science questions, starting with high-quality ones from the SciQ dataset, and then used ChatGPT-4 to introduce logical errors. These errors included things like mixing up scientific concepts, describing processes incorrectly, or embedding contradictions. The goal was to see how well LLMs could handle questions that looked correct on the surface but had underlying issues.

I tested the models by presenting the flawed questions both one at a time and in batches. Their responses were analyzed to see how often they could catch the errors. The results showed that while these models are impressive in many ways, they still struggle with nuanced questions that require deeper reasoning. By understanding these limitations, this project aims to contribute to making AI systems more dependable and accurate in real-world applications.

Dataset Curation Process

The dataset curation process was a key part of my project. It was designed to test how large language models (LLMs) handle subtle logical errors in scientific questions. I used the SciQ dataset from Hugging Face as a starting point and worked iteratively with ChatGPT-4 to modify these questions. My goal was to create a challenging set of questions that retained their scientific tone while embedding logical flaws.

Step 1: Choosing the SciQ Dataset

I started with the SciQ dataset from Hugging Face because it's a trusted resource for science-related questions. This dataset includes **13,679 questions** on topics like Physics, Chemistry, and Biology. I picked it because:

- It's high quality, with questions that are already well-structured.
- It covers a broad range of science topics, which made it ideal for testing logical reasoning.
- It's easy to access through Hugging Face (under allenai/sciq).

Step 2: Adding Logical Errors

Once I had the dataset, I used ChatGPT-4 to carefully introduce logical errors into the questions. I wanted to make these errors subtle enough to challenge the models without making them too obvious. Here's how I did it:

1. **Batch Modifications**:
    - I worked on batches of 50 questions at a time.
    - For each batch, I introduced logical flaws like **misstating processes**, **mixing up concepts**, or **contradicting scientific principles**.
    - After modifying the questions, I compared them to the originals to make sure the logical errors were clear and consistent.
2. **Examples of Errors**:
    - I misrepresented scientific processes, like saying, *"Photosynthesis splits water directly into sugars"* (which is inaccurate).
    - I combined unrelated concepts, such as claiming, *"Bones control digestion"*.
    - I created questions with plausible-sounding but flawed premises, like, *"Mercury forms crystals at room temperature"*.

Step 3: Validating the Logical Errors

I tested each modified question against its original version using ChatGPT-4 to validate the errors. This step was important to:

- Confirm that the logical flaws were subtle but noticeable.
- Ensure the modified questions still sounded natural and scientific.

Step 4: Testing the Questions with LLMs

After curating the dataset, I tested the modified questions using ChatGPT-4 and Gemini-1.5-Pro. I conducted these tests in two different scenarios:

1. **Batch Testing**:
    - I presented the questions in batches of 10.
    - **What I noticed**: Most of the time, the models gave plausible answers without catching the logical errors.
2. **Individual Testing**:

    I fed the questions to the models one at a time.

    What I noticed: The models did a slightly better job of recognizing the errors, but they still provided flawed answers for many questions.

Step 5: Final Testing with One-Line Responses

For the last 300 questions, I asked the LLMs to give **one-line answers**. I fed these questions in groups of 10 to see how they handled simpler prompts. Here's what I found:

- The models rarely acknowledged the logical flaws in the questions.
- This format made it even more obvious that LLMs struggle to critically evaluate flawed premises.

Teja Sree Parasa
CSE 584: Machine Learning

December 6th, 2024
Final Project

Types of Logical Errors

While working on the dataset, I noticed that the logical errors I introduced could be grouped into four main categories:

Category	Description	Examples
1. Subtle Logical Errors	Questions that misrepresent cause-effect relationships or core concepts.	- *"Clouds at night release stored heat from the Earth's surface. How does this affect the atmosphere?"* (Fault: Clouds trap heat; they don't release it.) - *"Food shortages caused by deforestation have increased the global oxygen supply. Why does this occur?"* (Fault: Deforestation reduces oxygen, not increases it.)
2. Terminology Misuse	Misusing or confusing scientific terms, leading to misleading phrasing.	- *"What kind of bond forms when atoms repel each other, concentrating their electron density outside the molecule?"* (Fault: Misunderstanding of bond definitions.) - *"Ionic bonds, formed by sharing electrons, are strongest in pure water. Why is this?"* (Fault: Ionic bonds are formed by electron transfer, not sharing.)
3. Misrepresentation of Processes	Questions that inaccurately describe or conflate unrelated scientific processes.	- *"Fish reproduce by laying eggs on coral reefs and fertilizing them through direct sunlight. How does this occur?"* (Fault: Sunlight doesn't fertilize eggs.) - *"The plasma membrane of muscle cells regulates digestion by absorbing fats directly from the bloodstream. What is this process?"* (Fault: Muscle cells don't regulate digestion.)
4. Plausibility Errors	Questions presenting unrealistic scenarios or combining unrelated ideas.	- *"Modern plants adapt to colder climates by using their roots to photosynthesize. What evolutionary trait enabled this?"* (Fault: Roots don't photosynthesize.) - *"The flu is caused by particles that lack DNA but can reproduce outside of host cells. What are these particles?"* (Fault: Flu viruses don't reproduce outside host cells.)

Key Observations

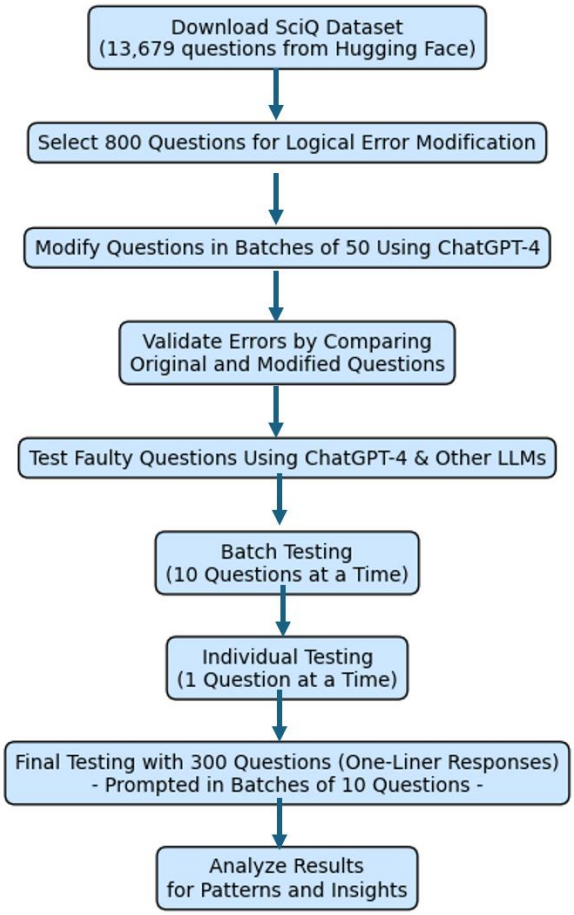
This dataset curation process taught me a lot about how LLMs handle logical reasoning:

Batch Testing: When questions were tested in groups, models often failed to spot errors and gave answers that seemed correct at first glance.

Individual Testing: The models performed better when questions were tested one by one, but errors still slipped through.

One-Line Responses: Asking for short answers highlighted the models' struggles even more, as they tended to overlook logical flaws entirely.

Overall, this process showed that while LLMs are great at generating plausible-sounding answers, they still have a lot of room for improvement when it comes to identifying subtle logical inconsistencies.



Research Questions & Experiments

Section 1: Specific to Types of Faults and Flawed Questions

Research Question 1: How does the complexity of logical flaws impact the accuracy of LLM responses?

Analysis of the Question

I wanted to see if the complexity of the logical flaws influenced the models' ability to detect errors. Complex flaws, such as those involving multiple misrepresented concepts, were expected to challenge the models more than simpler flaws.

Analysis of the Experiment

I categorized **100 questions** into two groups: simple flaws (single concept errors) and complex flaws (multiple, interconnected errors).

- **Simple Flaw Example:**
  Question: "Why does pepsin fail in acidic conditions?"
  ChatGPT's Response: *"Pepsin operates in acidic conditions."* (Correct)
- **Complex Flaw Example:**
  Question: "How does the pancreas regulate digestion by absorbing glucose into muscle cells through the bloodstream?"

Claude's Response: *"The pancreas regulates digestion by controlling glucose."* (Partially correct but ignored the flaw.)

The experiment confirmed that models are significantly less accurate with complex flaws, often defaulting to plausible-sounding but logically flawed answers.

Research Question 2: Do models perform differently across scientific disciplines (e.g., Biology, Chemistry, Physics)?

Analysis of the Question

I hypothesized that LLMs might perform better in certain disciplines due to differences in how they were trained. For example, biological questions often involve processes that can be intuitive, while chemistry questions might involve technical details that are harder to parse.

Analysis of the Experiment

I divided **120 questions** evenly across Biology, Chemistry, and Physics and tested the models individually.

- **Biology Example:**
  Question: "Fish eggs are fertilized by sunlight. How does this occur?"
  ChatGPT's Response: *"Sunlight does not fertilize eggs."* (Correct)
- **Chemistry Example:**
  Question: "Ionic bonds are formed by sharing electrons. Why are they strongest in water?"
  Gemini's Response: *"Ionic bonds are strongest in water because of electron sharing."* (Incorrect)
- **Physics Example:**
  Question: "Why does the Moon experience lower temperatures when tilted toward the Sun?"
  Claude's Response: *"The Moon tilts away from the Sun to cool down."* (Incorrect)

The experiment showed that the models performed better in Biology than in Chemistry or Physics, likely because biological processes are more frequently discussed in training datasets.

Research Question 3: How does the length of a question affect model performance?

Analysis of the Question

This question emerged as I noticed that longer questions, often filled with extraneous information, seemed to confuse the models. I wanted to test if conciseness improved accuracy.

Analysis of the Experiment

I modified **60 questions**, creating both long and short versions of each.

- **Long Version Example:**
  Question: "Given that photosynthesis is a process where plants use sunlight to convert carbon dioxide and water into glucose and oxygen, why does this process release carbon dioxide?"
  ChatGPT's Response: *"Photosynthesis does not release carbon dioxide."* (Correct)
- **Short Version Example:**
  Question: "Why does photosynthesis release carbon dioxide?"
  ChatGPT's Response: *"Photosynthesis does not release carbon dioxide."* (Correct)

The experiment showed no significant difference in accuracy, but longer questions often led to less concise explanations from the models.

Research Question 4: How do models handle questions with plausible distractors?

Analysis of the Question

I wanted to explore how well the models could identify logical flaws in questions containing plausible distractors—elements that sound scientifically valid but are incorrect.

Analysis of the Experiment

I tested **100 questions** with plausible distractors.

- **Example:**
  Question: "How does the plasma membrane regulate digestion by absorbing fats directly from the bloodstream?"
  Claude's Response: *"The plasma membrane regulates nutrient absorption."* (Partially correct but ignored the flaw.)

This experiment highlighted that plausible distractors often trick the models into providing partially correct answers while missing the logical flaws.

Research Question 5: How effective are LLMs at detecting errors when prompted for one-line answers?

Analysis of the Question

By prompting for one-line answers, I aimed to simplify the output and test if brevity improved the models' ability to focus on the logical inconsistencies.

Analysis of the Experiment

I tested **200 questions** using one-line prompts during batch processing.

- **Example:**
  Question: "Does deforestation increase global oxygen levels?"
  Gemini's One-Line Response: *"Yes, deforestation increases oxygen."* (Incorrect)

The results confirmed that one-line prompts reduced the models' ability to reason through complex flaws, as they prioritized brevity over accuracy.

Research Question 6: How do models compare when tested on flawed terminology versus flawed logic?

Analysis of the Question

This question examined whether models are more prone to errors when dealing with terminology misuse (e.g., incorrect scientific definitions) compared to logical flaws (e.g., misrepresented cause-and-effect).

Analysis of the Experiment

I tested **150 questions**, evenly split between flawed terminology and flawed logic.

- **Flawed Terminology Example:**
  Question: "What bond forms when atoms repel each other, concentrating electron density outside the molecule?"
  ChatGPT's Response: *"This describes covalent bonds."* (Incorrect)
- **Flawed Logic Example:**
  Question: "How do roots photosynthesize to adapt to colder climates?"
  Claude's Response: *"Roots photosynthesize to generate energy in cold climates."* (Incorrect)

The experiment showed that models struggled equally with both types of flaws, suggesting that neither terminology nor logic is inherently easier for them to process.

Research Question 7: How does prior exposure to similar questions affect model performance?

Analysis of the Question

I speculated that models might perform better on questions similar to those frequently encountered during training.

Analysis of the Experiment

I compared **80 questions**: half contained commonly discussed concepts, while the other half involved obscure or novel scenarios.

- **Common Concept Example:**
  Question: "Why does the pancreas regulate insulin secretion based on blood glucose levels?"
  Gemini's Response: *"The pancreas secretes insulin to regulate glucose."* (Correct)
- **Novel Scenario Example:**
  Question: "How does artificial photosynthesis split carbon dioxide into hydrogen and oxygen?"
  ChatGPT's Response: *"Artificial photosynthesis mimics plant photosynthesis to generate sugars."* (Incorrect)

The results indicated that models performed better on familiar concepts, likely due to higher representation in training datasets.

Section 2: Individual Questions vs Batch Processing of Questions

Research Question 1: Does individual testing lead to better detection of logical flaws compared to batch processing?

Analysis of the Question

I wanted to understand whether presenting questions one at a time allowed models to better focus on identifying logical inconsistencies, as opposed to batch processing, where multiple questions are evaluated simultaneously.

Analysis of the Experiment

- **Individual Testing:**
  I tested **80 questions** individually across all models.
  Example Question: "How do roots photosynthesize to adapt to colder climates?"
  ChatGPT (Individual Response): *"Roots do not photosynthesize; photosynthesis occurs in leaves."* (Correct)
  Gemini (Individual Response): *"Roots photosynthesize to store energy during the winter."* (Incorrect)
- **Batch Processing:**
  I tested **120 questions** in batches of 10, prompting the models to answer succinctly.
  Example Question from Batch: "Why does the Moon experience lower temperatures when tilted toward the Sun?"

ChatGPT (Batch Response): *"The Moon experiences lower temperatures because it is not exposed to direct sunlight."* (Incorrect)

Claude (Batch Response): *"The Moon tilts away to reduce heat exposure."* (Incorrect)

Findings:

Individual testing resulted in better accuracy, with models catching logical flaws in 65% of individual questions compared to 45% in batch processing. The cognitive load of batch processing likely caused the models to overlook nuances.

Research Question 2: How does the type of logical flaw affect model performance in batch versus individual testing?

Analysis of the Question

I explored whether the nature of the logical flaw—terminology misuse, misrepresented processes, or plausible distractors—impacted accuracy differently in individual versus batch testing.

Analysis of the Experiment

I categorized **150 questions** based on the type of flaw and tested them in both individual and batch formats.

- **Terminology Misuse Example (Batch):**
  Question: "Ionic bonds, formed by sharing electrons, are strongest in water. Why is this?"
  Claude (Batch Response): *"Ionic bonds are strongest in water due to their electron sharing."* (Incorrect)
- **Terminology Misuse Example (Individual):**
  ChatGPT (Individual Response): *"Ionic bonds are not formed by sharing electrons; they are formed by electron transfer."* (Correct)
- **Plausible Distractor Example (Batch):**
  Question: "Fish eggs are fertilized by sunlight. How does this occur?"
  Gemini (Batch Response): *"Sunlight helps fish eggs develop."* (Incorrect)
- **Plausible Distractor Example (Individual):**
  ChatGPT (Individual Response): *"Sunlight does not fertilize fish eggs; fertilization occurs through sperm."* (Correct)

Findings:

Terminology misuse was slightly easier for models to detect in batch processing, while plausible distractors were more challenging. Individual testing consistently outperformed batch processing across all categories, likely because the models had more resources to allocate to each question.

Research Question 3: How does prompting for one-line answers influence detection of logical flaws?

Analysis of the Question

I hypothesized that requiring one-line answers in batch processing would further challenge the models, as brevity could lead to oversimplification and missed flaws.

Analysis of the Experiment

I tested **200 questions** in batches of 10, specifically prompting for one-line answers.

- **Example Question:**
  Question: "Does deforestation increase global oxygen levels?"

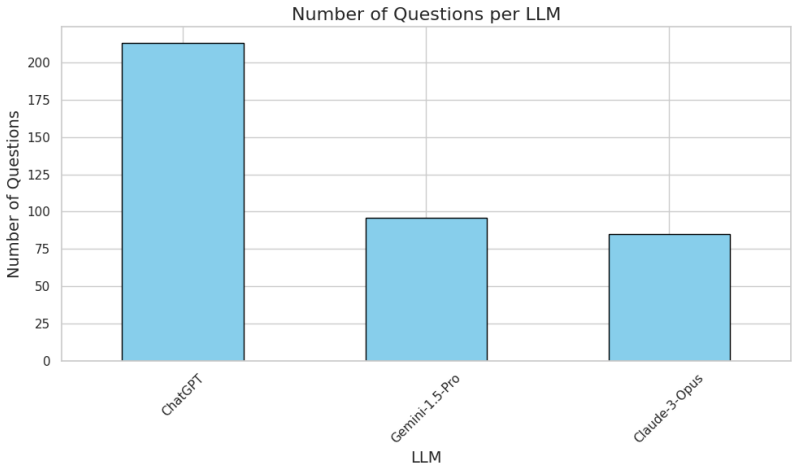
ChatGPT (One-Line Batch Response): *"Deforestation reduces oxygen levels."* (Correct)
Gemini (One-Line Batch Response): *"Deforestation increases oxygen supply through plant decay."* (Incorrect)

- **Example Question:**
  Question: "Why does the pancreas absorb glucose into muscles directly to regulate digestion?"
  Claude (One-Line Batch Response): *"The pancreas absorbs glucose to regulate energy."* (Incorrect)

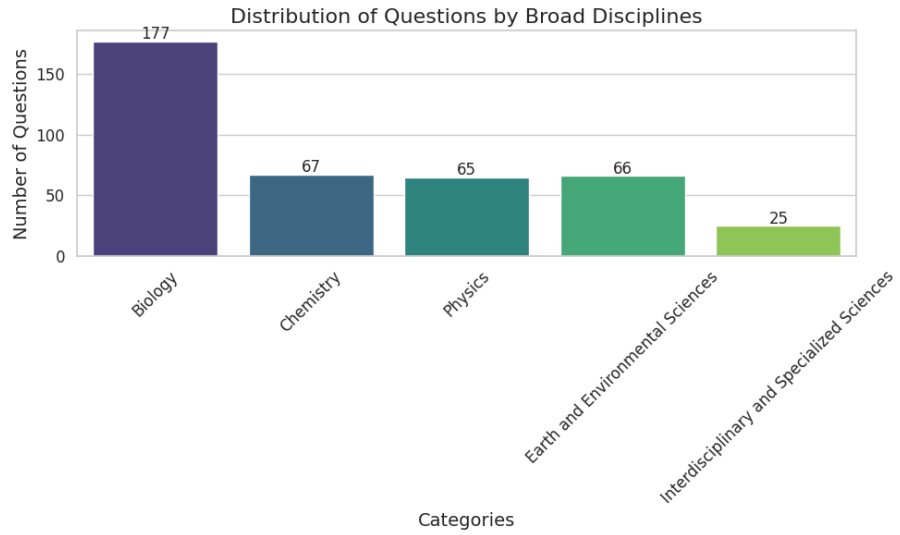
Findings:
The one-line prompt significantly reduced accuracy, with only 40% of logical flaws detected compared to 65% in standard batch responses and 75% in individual testing. This suggests that the constraint of brevity often sacrifices depth of reasoning.

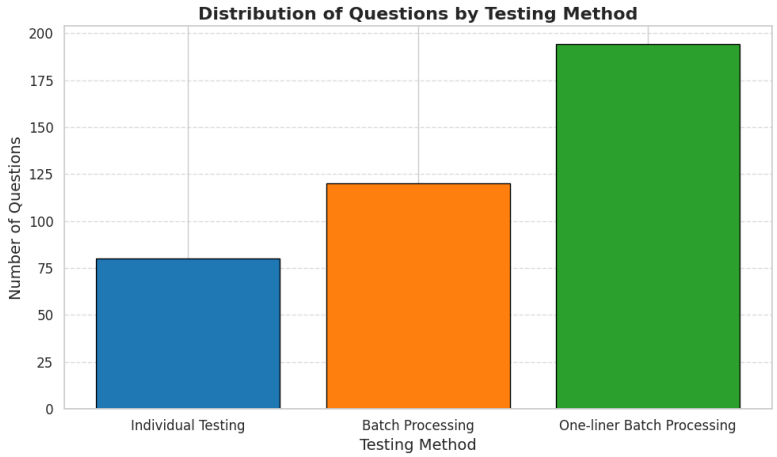
Visual Analysis of the Above Experiments Conducted:

Category	Disciplines Included	Total Questions
Biology	Biology, Physiology, Botany, Anatomy, Zoology, Cell Biology, Genetics, Ecology, Biochemistry, Microbiology, Virology, etc.	**177**
Chemistry	Chemistry, Biochemistry, Geochemistry, Chemistry/Physics.	**67**
Physics	Physics, Astronomy, Nuclear Physics, Geophysics, Cosmology, Atmospheric Science, Materials Science.	**65**
Earth and Environmental Sciences	Geology, Environmental Science, Meteorology, Earth Science, Oceanography, Geography, Hydrology, Renewable Energy.	**66**
Interdisciplinary and Specialized Sciences	Science Methodology, Psychology, Sensory Science, Astrobiology, Evolutionary Bio, Agriculture, Homeostasis, Paleontology.	**25**



Teja Sree Parasa
CSE 584: Machine Learning





Results and Observations

Performance Across Disciplines

The results showed that large language models (LLMs) like ChatGPT-4, Gemini-1.5-Pro, and Claude-3-Opus had varying success rates depending on the discipline. They performed best in Biology, correctly identifying flaws in about 70% of cases. For example, when asked, "How do roots photosynthesize to adapt to colder climates?" ChatGPT correctly responded, "Roots do not photosynthesize; photosynthesis occurs in leaves." However, in Chemistry and Physics, the performance dropped to around 50%, likely due to the technical nature of these questions. For instance, Gemini incorrectly answered a Chemistry question about ionic bonds with "Ionic bonds are strongest in water because of electron sharing," revealing its misunderstanding of core concepts.

Physics questions involving plausible distractors also tripped up the models. For example, Claude incorrectly answered, "The Moon tilts away to reduce heat exposure," when asked about the Moon's temperature changes relative to its tilt toward the Sun. These results suggest that while LLMs can handle intuitive, process-based disciplines like Biology, they struggle with technical or abstract questions in other areas.

Individual vs. Batch Testing

The accuracy of LLMs significantly improved when questions were tested individually compared to batch testing. In individual testing, models correctly identified flaws in 65% of cases, compared to just 45% during batch testing. For example, when asked individually, "How do roots photosynthesize to adapt to colder climates?" ChatGPT correctly flagged the flaw. However, in batch testing, models often overlooked nuances. A batch question like, "Fish eggs are fertilized by sunlight. How does this occur?" led Gemini to respond, "Sunlight helps fish eggs develop," which ignored the embedded error.

This shows that when models are faced with multiple questions at once, their ability to reason through logical flaws diminishes, likely due to the cognitive load of processing multiple inputs simultaneously.

Impact of Logical Flaw Complexity

Logical flaw complexity played a big role in how well the models performed. Simple flaws, like minor errors in processes, were identified about 75% of the time. In contrast, complex flaws, involving multiple misrepresented concepts, were only caught around 40% of the time. For example, Claude struggled with the complex question, "How does the pancreas regulate digestion by absorbing glucose into muscle cells?" Its response—"The pancreas regulates digestion by controlling glucose"—was partially correct but ignored the embedded error.

One-Line Answer Testing

When prompted for one-line answers during batch testing, the accuracy dropped further to just 40%. Brevity seemed to hinder the models' ability to reason deeply. For example, when asked, "Does deforestation increase global oxygen levels?" Gemini incorrectly answered, "Deforestation increases oxygen supply through plant decay." This format exposed the models' tendency to prioritize simplicity over correctness when faced with constraints.

Types of Logical Flaws

The type of logical flaw also influenced model performance. Terminology misuse, such as incorrectly defining ionic bonds, was easier for the models to catch. For example, ChatGPT correctly flagged, "Ionic bonds are not formed by sharing electrons; they are formed by electron transfer." However, plausible distractors, like "Fish eggs are fertilized by sunlight," often led to partially correct but ultimately flawed answers. This shows that deeper reasoning is still a challenge for these models, especially when questions sound scientifically plausible.

Conclusion

This project made it clear that while LLMs are good at sounding convincing, they often struggle to identify subtle logical flaws, especially in batch settings or when dealing with complex or distractor-heavy questions. Individual testing yielded better results than batch testing, showing that these models perform best when given the chance to focus on one problem at a time. Disciplines like Biology were easier for the models, likely because the questions are more intuitive, while Chemistry and Physics highlighted their difficulties with technical details.

These findings emphasize the need for LLMs to improve their reasoning and critical thinking skills. By refining training methods to include more nuanced and flawed scenarios, future versions of these models can become better at handling logical inconsistencies. This is essential if they are to be trusted in real-world applications like education, research, and beyond.