

CSE584: Machine Learning: Tools and Algorithms

Final Course Research Project Report

I collected a dataset of 120 questions in 4 disciplines Maths, Biology, Chemistry, Physics. I created question and modified the dataset which are available online by making minor changes to it by changing some logic. All questions were put into llama, gpt4o, gpt4, gpt4o1, Claude 3.5 and tested whether they give correct answer or not. Whichever questions the error wasn't detected, and LLM gave answer for the question made it into the dataset. I made sure that the data was spread around numerical, theoretical, ambiguous, numeric errors, wrong premise topics so that I can work on experimenting on the results provided in a wide range of domains to study the strengths and weaknesses. I chose GPT 4o1 to test and experiment on the dataset as it is one of the most advanced LLM. Based on the dataset I have designed Research Question and conducted Experiments to solve this research question and draw conclusion to these questions.

Research Questions:

- 1. Which LLM performed the most inaccurate solution for the questions?**
- 2. Why do GPT-4o, Gemini 1 Pro, and Claude 3.5 fail to detect any faultiness in the 120 questions, while O1 detects them and state how many they detect?**
- 3. Are there patterns in the questions flagged as faulty by O1 that explain why they align with other models?**
- 4. What if I ask O1 to detect faults and what if I ask O1 to answer question. Will the values match. It should match because if it is faulty then they should not provide an answer and show**
- 5. What type of discipline does is o1 struggling out figure out faulty or not?**
- 6. What type of discipline was harder to make dataset of?**
- 7. How many detected were Numerical and Theoretical faults were detected?**
- 8. Categorize the question type which were determined as faults. Ambiguous , wrong premise, Numeric faults.**

Experiment 1:

All these question I put through GPT o1 because o1 is the most advance LLM model in the market which can answer any question accurately and compare its answer with used LLM for dataset creation and check whether it gives proper answer than LLM. GPT-O1 is recognized for its state-of-the-art architecture, capable of understanding and generating human-like responses across a wide range of domains and complexities.

After conducting this experiment I saw that some of the answers were still the same as I have conducted in dataset.

- Question 1:

A nerve impulse traveling at 50 meters per second takes 2 seconds to travel along a neuron that is 100 meters long. Is this correct?

Answer in Dataset :

Yes, that's correct.

We can use the formula:

Time = Distance / Speed \rightarrow

Given:

Distance = 100 meters

Speed = 50 meters/second

Calculating time:

Time = 100 meters / 50 meters/second = 2 seconds

Therefore, it would take 2 seconds for a nerve impulse traveling at 50 meters per second to travel along a 100-meter-long neuron.

Answer given by O1:

Mathematically, yes: $50 \text{ m/s} * 2 \text{ s} = 100 \text{ m}$. (Though neurons this long don't occur in humans.)

Reason why this question was wrong:

Path of impulse is not said. Electrical impulse could just stop because of external factor and

nothing is said about it. Start and end point is not mentioned. Question is ambiguous and can be interpreted in multiple ways.

- **Question 2:**

A man steps over 0.5m hurdle. Minimum height he should jumped?

Answer in dataset:

If a man steps over a 0.5-meter hurdle, the minimum height he could have jumped is just enough to clear the hurdle, which would be slightly more than 0.5 meters.

Answer given by O1:

Slightly more than 0.5 m to clear the hurdle.

Reason why this question is faulty:

He steps over hurdle no need to jump.

- **Question 3:**

A train travels 100 miles at 50 mph. How long does it take?

Answer in dataset:

If a man steps over a 0.5-meter hurdle, the minimum height he could have jumped is just enough to clear the hurdle, which would be slightly more than 0.5 meters.

Answer given by O1:

Time=distance/speed=100/50=2 hours.

Reason why this question is faulty:

He steps over hurdle no need to jump.

- **Question 4:**

A 5 kg object moving at 10 m/s has energy of 500 J. True or false?

Answer in dataset:

"False. The kinetic energy calculation for a 5 kg object moving at 10 m/s is 250 J, not 500 J as stated in the query.

The kinetic energy formula

$$KE = \frac{1}{2}mv^2$$

confirms the calculation:

$$\text{Mass (m)} = 5 \text{ kg}$$

$$\text{Velocity (v)} = 10 \text{ m/s}$$

$$\text{Kinetic Energy} = \frac{1}{2} \times 5 \times (10)^2 = 250 \text{ J}$$

Answer given by O1:

$$KE = \frac{1}{2}mv^2 = \frac{1}{2} \times 5 \times 100 = 250 \text{ J, not 500 J. False.}$$

Reason why this question is faulty:

Nothing is given about kinetic but it assumes kinetic energy

Some of the answers were changed and it guessed correctly

- Question 1:

Mendel's Law of Segregation states that alleles of different genes assort independently of one another.

Based on the Law of Segregation, predict the phenotypic ratio in a dihybrid cross.

Answer in dataset:

Mendel's Law of Segregation

Key Principles

During meiosis, each allele has an equal and random chance of being selected and passed on to offspring

Alleles segregate independently during gamete formation

Each gamete receives only one gene copy, selected randomly

Dihybrid Cross Prediction

In a dihybrid cross involving two traits, Mendel discovered a specific phenotypic ratio of 9:3:3:1. This ratio is derived from the independent assortment of alleles:

9/16: Double dominant phenotype

3/16: First dominant, second recessive phenotype

3/16: First recessive, second dominant phenotype

1/16: Double recessive phenotype

Mathematical Derivation

The ratio is calculated using the product rule: $\frac{3}{4} \times \frac{3}{4} = \frac{9}{16}$

$$\frac{4}{3} \times \frac{4}{3} = \frac{16}{9}$$

Assumptions for Ratio

Genes assort independently

One allele at each locus is completely dominant

Phenotypes can be distinguished unambiguously

No gene interaction occurs

Example Scenario

In a cross of pea plants with different seed traits (colour and shape), the 9:3:3:1 ratio emerges when crossing two heterozygous parents.

Answer given by O1:

Mendel's Law of Segregation states that allele pairs separate during gamete formation.

Mendel's Law of Independent Assortment states that alleles of different genes assort independently.

Reason why this question is faulty:

That's the Law of Independent Assortment. The question confuses two laws

- **Question2:**

A guy ate fruit in weight to time ratio of 1:-3 and girl ate fruit in ratio of 3:-2 who is faster per pound

Answer in dataset:

To determine who ate faster per pound of fruit, we can analyse the ratios provided, adjusting them to a common format for clearer comparison. The ratios given are in terms of weight to time (pounds per minute, assuming time is in minutes), where negative time indicates the amount of time it takes to eat a certain weight.

Understanding the Ratios:

Guy's Ratio: 1:-3 implies it takes him 3 minutes to eat 1 pound of fruit.

Girl's Ratio: 3:-2 implies it takes her 2 minutes to eat 3 pounds of fruit.

Converting to Pounds per Minute:

Guy: Rate = 1 pound / 3 minutes = 0.333 pound per minute

Girl: Rate = 3 pound / 2 minutes = 1.5 pound per minute

Comparison:

The girl's eating rate is 1.5 pounds per minute.

The guy's eating rate is 0.33 pounds per minute.

Conclusion:

The girl ate faster per pound, as her rate of eating (1.5 pounds per minute) is greater than the guy's rate (0.33 pounds per minute).

Answer given by O1:

Negative ratio nonsensical. Faulty.

Reason why this question is faulty:

Ratio can't be negative for rate of eating

- **Question 3:**

A cube has a volume of -64 cm^3 . What is the length of one edge? Is it 4 cm? Just give me the explanation for value.

Answer in dataset:

Yes, in this case, the length of one edge of the cube is -4 cm.

Here's why:

Volume of a cube = edge^3

So, if the volume is -64 cm^3 , we can set up the equation:

$\text{edge}^3 = -64$

To find the edge length, we need to find the cube root of -64.

The cube root of -64 is -4.

Therefore, the length of one edge of the cube is -4 cm.

Answer given by O1:

Negative volume not possible. Faulty.

Reason why this question is faulty:

Volume can't be negative

- **Question 4:**

The angle covered by the hands of a clock for going from 15:00 to 21:00 is what when it crosses the time first time?

Answer in dataset:

From 15:00 to 21:00, the hour hand moves 6 hours.

Since a full circle (360 degrees) is covered in 12 hours, the hour hand moves 30 degrees in one hour.

So, in 6 hours, it moves $6 * 30 = 180$ degrees.

Therefore, the angle covered by the hands of the clock is 180 degrees

Answer given by O1:

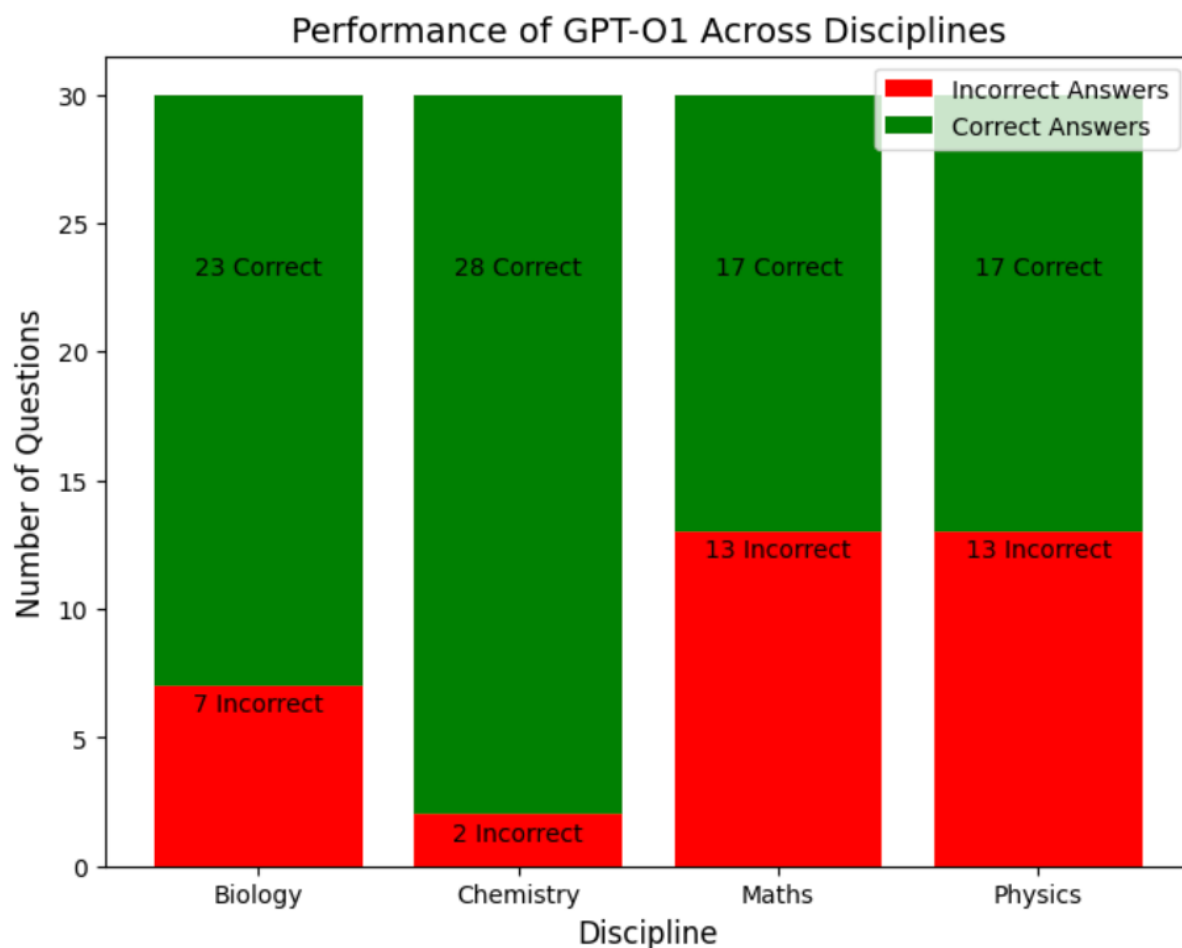
Ambiguous question. Faulty due to lack of clarity.

Reason why this question is faulty:

Never mentioned which hand angle we have to calculate.

Below is the visualization of analysis of answers given by o1:

Discipline	Correct Answers	Incorrect Answers	Total Questions	Percentage Correct	Percentage Incorrect
Biology	23	7	30	76.67%	23.33%
Chemistry	28	2	30	93.33%	6.67%
Maths	17	13	30	56.67%	43.33%
Physics	17	13	30	56.67%	43.33%



- Worst Performance which GPT-O1 performed the best in Chemistry, with only 2 incorrect answers.
- Best Performance which GPT-O1 struggled most in Maths and Physics, with 13 incorrect answers each.
- Judging by the responses so far given It seems GPT-O1 performs worse on conceptual or fact-based questions (e.g., Chemistry and Biology) and performs better with logical or numerical domains (e.g., Maths and Physics).

Experiment 2:

Out of 4 disciplines which is Maths, Biology, Physics, Chemistry, I used o1 to Identify which are faulty questions and correct them and give me count of faulty questions. Also give count per discipline how many did it count as faulty.

Questions which are not spotted as faulty

Question 1:

If hypotenuse of triangle is $4\sqrt{2}$ find legs

Answer by o1:

Possibly 4 and 4 for a 45-45-90. Not necessarily faulty.

Reason why it was faulty:

I have not given the type of triangle. It went for the most obvious isosceles triangle and gave the output.

Question 2:

A line segment which touches the circle at exactly one point

Answer by o1:

Tangent. Correct. Not faulty.

Reason why it was faulty:

It misinterpreted line segment as line. Tangent is a line which touches a circle. Tangent is not a line segment.

Question 3:

Sum of angles in a triangle is 360° .

Answer by o1:

Sum is 180° .

Reason why it was faulty:

I never mentioned whether we have to calculate interior angle sum or exterior angle sum. The question is ambiguous and given information is lacking the specific angle we need to consider.

Question 4:

I am talking about 3d shapes

I have got a rectangle base and height of 5 m

dimension of rectangle is 9X10

what is the volume

Answer by o1:

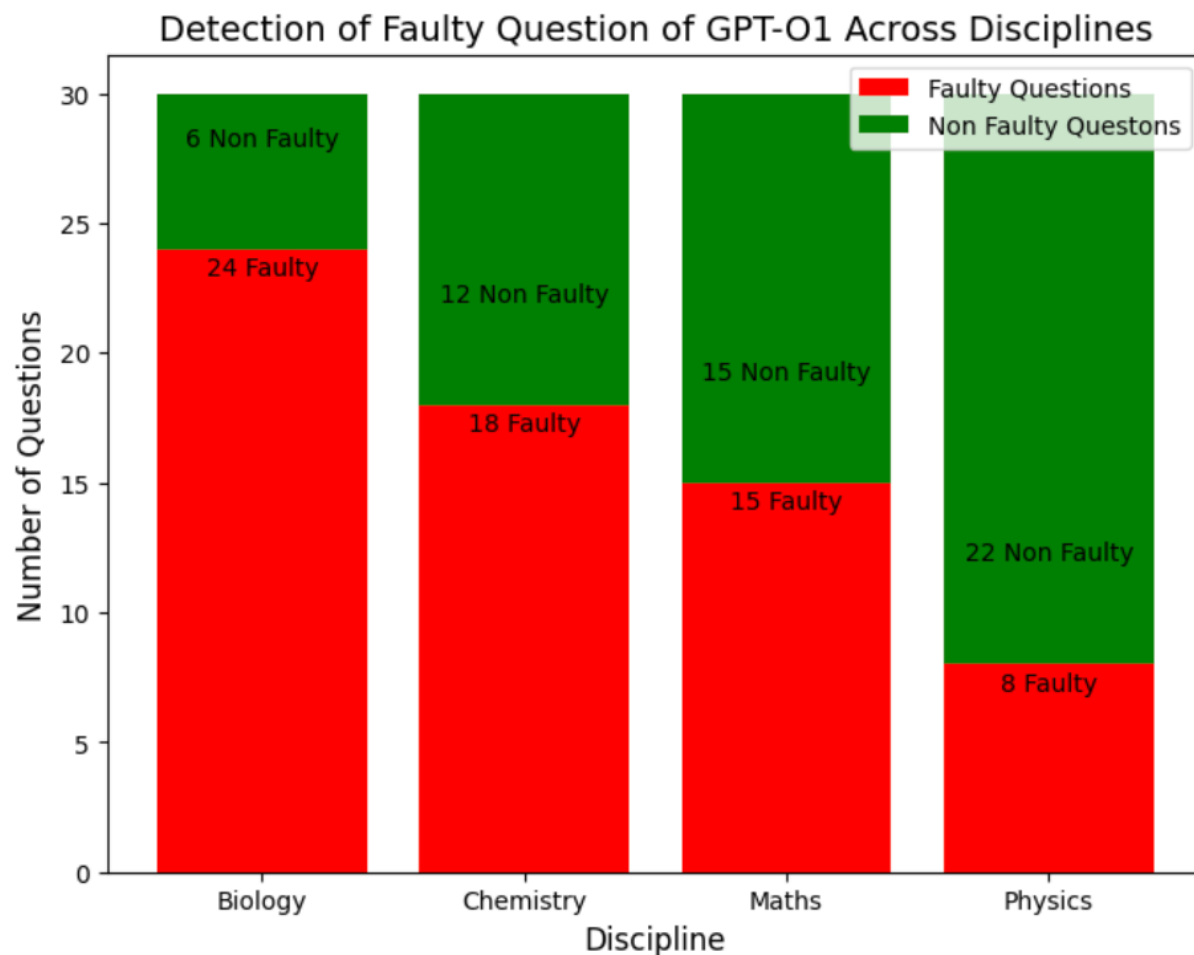
Volume=9x10x5=450 correct. Not faulty.

Reason why it was faulty:

I never mentioned in the question that shape was cuboid. It could have been a trapezium structured into 3D model. The question requires more information to calculate the volume.

Below is the visualization of analysis of answers given by o1:

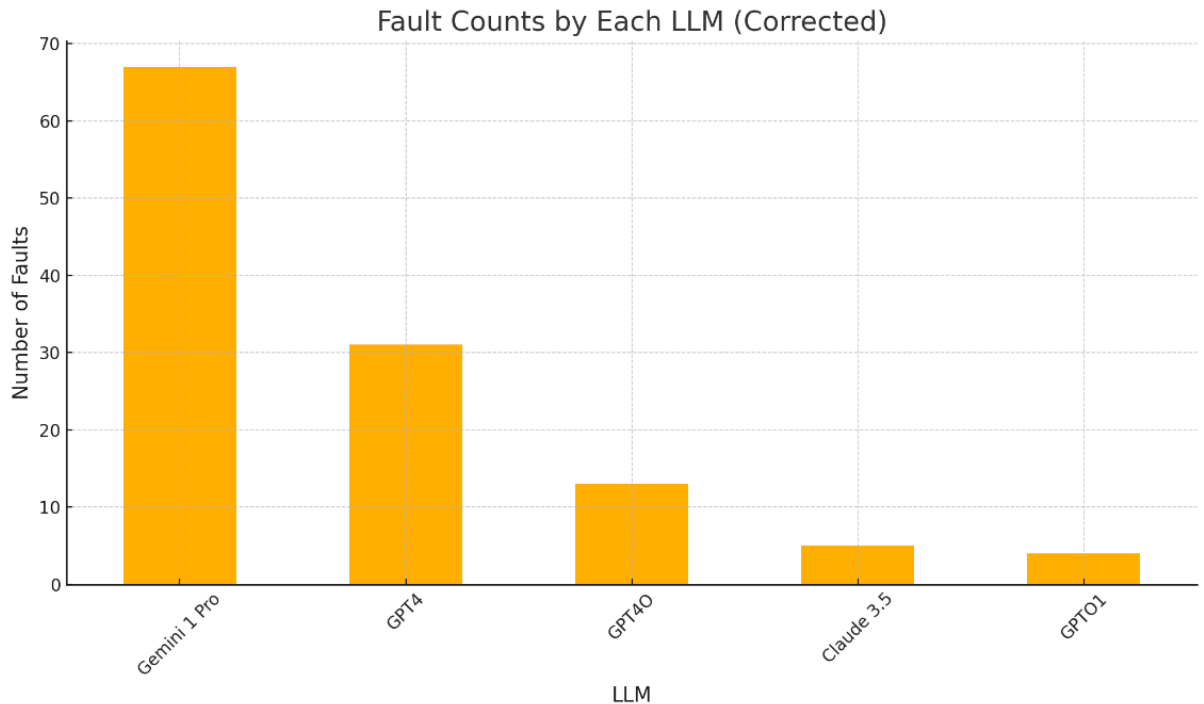
Discipline	Faulty Questions	Non-Faulty Questions	Total Questions	Percentage Faulty	Percentage Non-Faulty
Biology	24	6	30	80.00%	20.00%
Chemistry	18	12	30	60.00%	40.00%
Physics	15	15	30	50.00%	50.00%
Maths	8	22	30	26.67%	73.33%



- GPT-O1 performed **best in Biology**, detecting the highest number of faulty questions
- GPT-O1 struggled the most in **Physics**, detecting only 8 faulty questions and identifying 22 as non-faulty.

Research Questions:

1. Which LLM performed the most inaccurate solution for the questions?



As we can see Gemini 1 Pro has the highest number of question which the LLM answers incorrectly.

2. Why do GPT-4o, Gemini 1 Pro, and Claude 3.5 fail to detect any faultiness in the 120 questions, while O1 detects them and state how many they detect?

The main difference between o1 and other LLM could be that Open AI has claimed that o1 is trained on intensive data rather than extensive but went in deep with the understanding of complex topics. o1 demonstrates superior performance in, Advanced mathematics (83% on IMO qualifying exam), Complex scientific problem-solving, Multi-step reasoning tasks, Fault detection and question analysis.

The dataset is narrower in scope but it claims to have been excellent in detecting ambiguity and faultiness in question.

3. Are there patterns in the questions flagged as faulty by O1 that explain why they align with other models?

Other model likely provide an answer based on assumption rather than rejecting the idea or faults in question Whereas o1 is maybe flagging a question as faulty if it thinks that question can be interpreted in multiple ways. We can see this happen during the experiment I conducted. When question was “Angle covered by clock from

15:00 to 21:00 first time” It gave answer as “Ambiguous question. Possibly faulty due to lack of clarity.”

4. What if I ask O1 to detect faults and what if I ask O1 to answer question. Will the values match. It should match because if it is faulty then they should not provide an answer and show

I think this happened because LLM likes to play on prompts which are given by the user. When asked to "detect faults," O1 focussed on identifying ambiguity, logical inconsistencies, or missing information in the question. When I asked question to "answer the question," O1 will focus on generating a correct answer, even if the question is ambiguous or faulty. This forces the model to assumptions, which can lead to hallucinations.

This can be inferred from the experiments I conducted. O1 detect 65 questions as faulty but only provide 35 answers different from the other LLMs and were correct.

5. What type of discipline does is o1 struggling out figure out faulty or not?

Biology: O1 detected 80% of questions as faulty. This high rate suggests that O1 is easily detects ambiguities in this discipline, likely due to the theoretical and descriptive nature of Biology questions.

Chemistry: 60% of questions were flagged as faulty, but the correct answer rate (93.33%) was very high. This suggests that Chemistry questions, even when faulty, o1 is making obvious assumption which may not be true in some cases

Physics: 50% of questions were flagged as faulty, and the correct answer rate was 56.67%.

Mathematics: Only 26.67% of questions were flagged as faulty, but the correct answer rate was low (56.67%).

6. What type of discipline was harder to make dataset of?

When I was creating dataset the most hardest to find dataset was Biology. And most easiest was Mathematics.

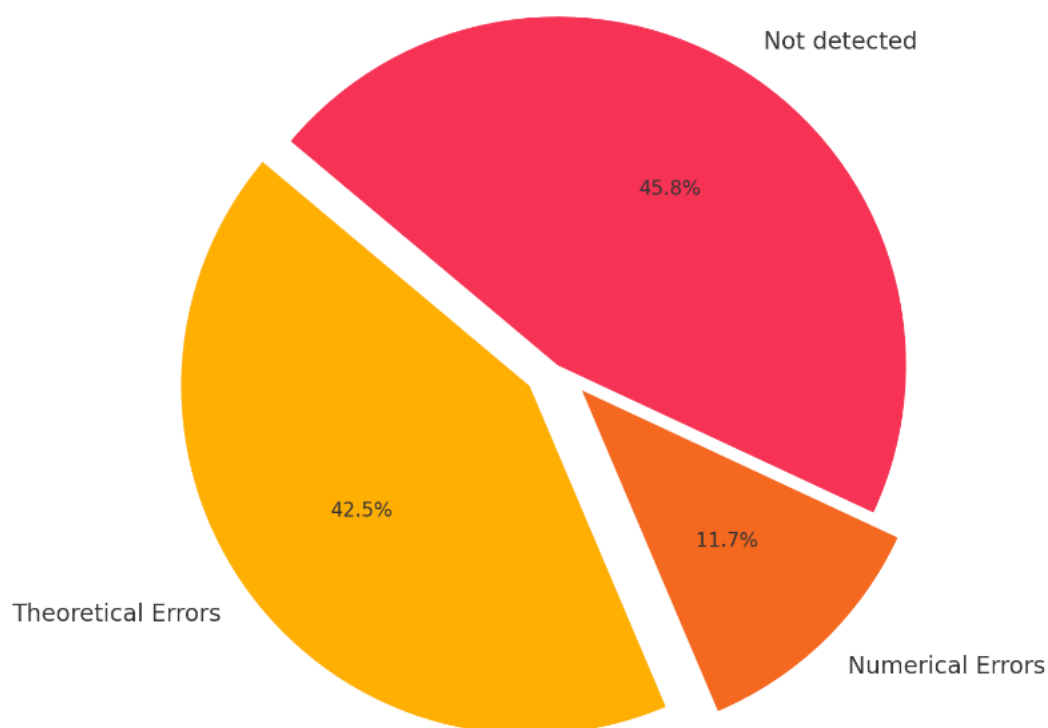
Biology as a subject is very theoretical and descriptive in nature. Biology involves detailed terminology and small errors are easily detected.

Unlike biology , math questions are less reliant on external context or real-world interpretation, reducing the complexity of creating plausible faulty questions. Many faulty math questions can be generated by simply altering existing problems according to known patterns of errors.

For example Let's say one question takes ratio as 1:-1 then we can increment -1 to -1000 and get 1000 faulty questions.

7. How many detected were Numerical and Theoretical faults were detected?

Distribution of Theoretical and Numerical Errors Out of 120 Questions



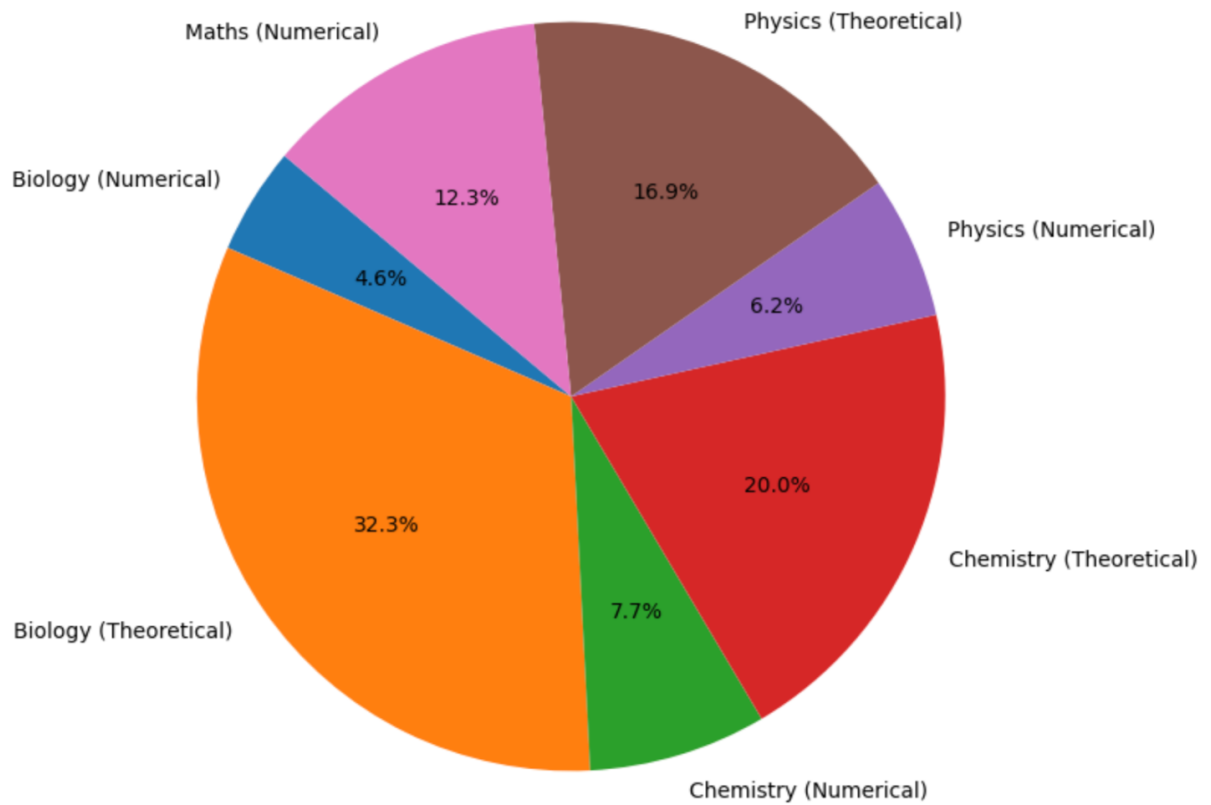
As previously analysed, theoretical and conceptual faults in Biology are challenging to miss because they can be easily verified using well-established knowledge. In contrast, numerical questions, while simpler to construct, are more difficult to detect as faulty, making them less likely to be flagged.

The count are as:

Numerical errors: 3 (Bio) + 5 (Chem) + 4 (Phys) + 2 (Math) = 14

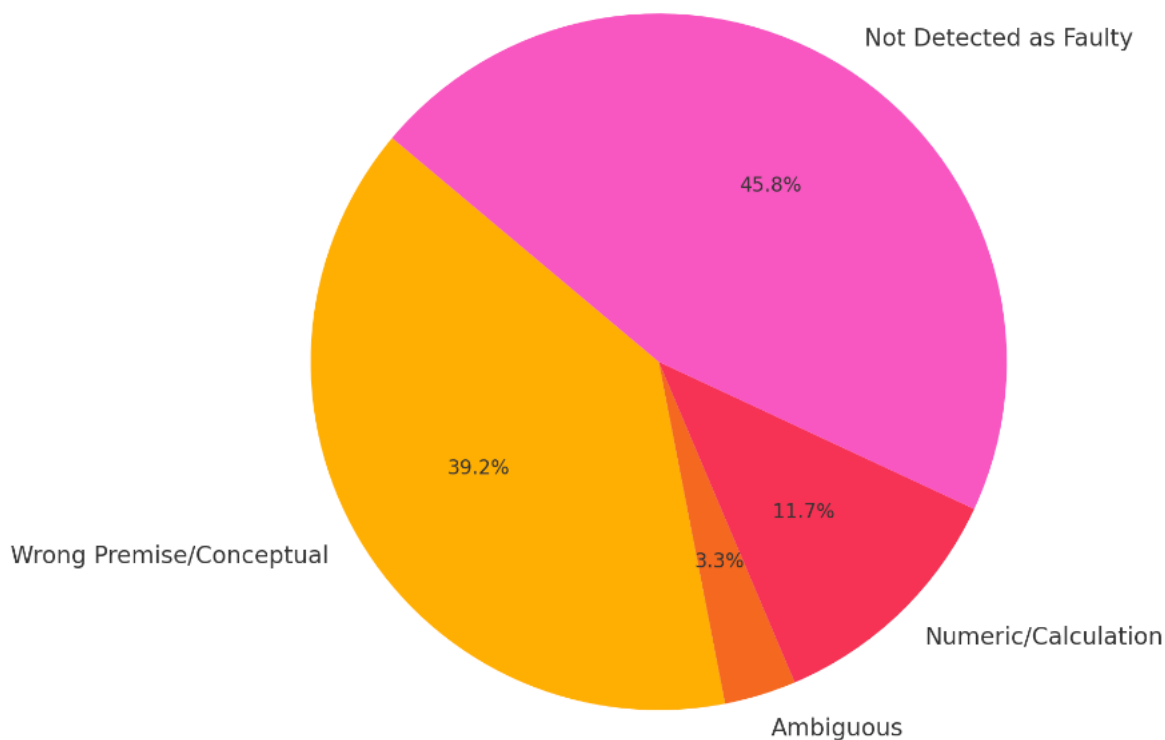
Theoretical errors: 21 (Bio) + 13 (Chem) + 11 (Phys) + 6 (Math) = 51

Proportion of Numerical and Theoretical Errors by Discipline



8. Categorize the question type which were determined as faults. Ambiguous , wrong premise, Numeric faults.

Proportion of Faulty and Non-Faulty Questions



Wrong premises are easy to detect because it is straightforward to verify whether the premise is correct or not. Ambiguous questions, however, are harder to identify because the LLM may not consider every possibility and might draw conclusions too early, even if information is missing. To answer the question, it may make obvious assumptions that are not mentioned, which could lead to an incorrect output

The count is as follows:

Wrong Premise: Biology (21) + Chemistry (11) + Physics (11) + Math (4) = 47

Ambiguous: Biology (0) + Chemistry (2) + Physics (0) + Math (2) = 4

Numeric/Calculation: Biology (3) + Chemistry (5) + Physics (4) + Math (2) = 14

Proportion of Errors by Discipline and Category

