CSE 584- Final Project Report

Varunsai Alaparthi
917181259(VSA5067)

Introduction:

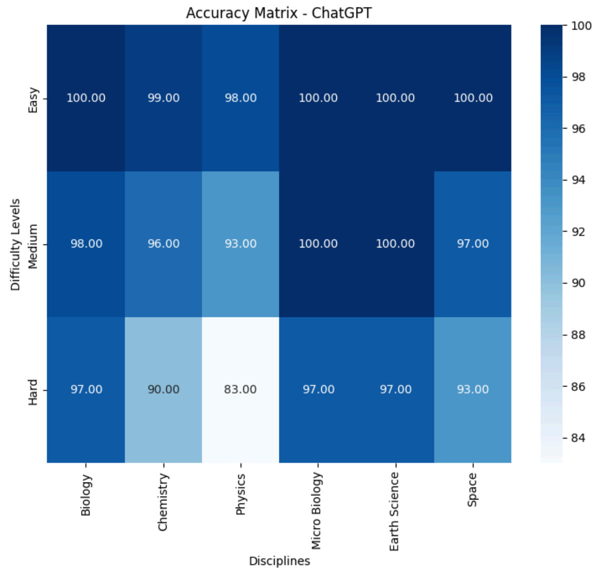
Created a dataset of faulty science questions from various disciplines like microbiology, Astronomy, Biology , Physics, chemistry , Earth science, zoology, History based science. I have ensured that the question is faulty in some sense, either conceptually or lacking some information or isn't practical. I have analysed different LLMs like chat-gpt, Claude, Gemini. I present some research questions and the respective analysis here.

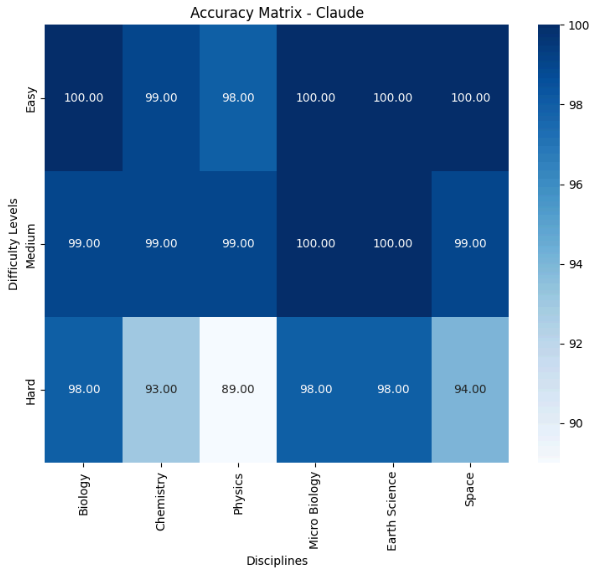
Here are some questions I tried to analyse :

Research Question-1 :

Which discipline was the toughest for the LLM's ?

Based on multiple experiments in different disciplines like microbiology, Astronomy, Biology , Physics, chemistry , Earth science, zoology, History based science, The least accuracy and most confusion was in physics. In other disciplines like for example Biology out of 100 faulty questions i tried out all the LLms are pointing out the mistakes/inaccuracies in the questions, Most top LLm have the fact checking mechanism where they do online search before answering which is making the task of fooling these models even more difficult especially for biology, zoology where most knowledge is available online. I have also tried with settings of turning off the net at least from my end , But I have seen mostly the same answers with exception of 2 which I have captured in my dataset.





The Difficulty levels are not accurate, I have decided them based on the level of the exams I have taken them from especially for physics, chemistry and space. For space , I have used dome confusing concepts from theory of relativity, and for physics and chemistry I have taken some tough questions from exams like JEE_Advanced , Olympiads where the level of toughness is a lot. In the rest of the fields I just gave the hardness based on how recent the concept/discovery is and also based on the length of my question as the model needs to then focus on the longer prompt. For Claude it’s not getting intimidated by the hard physics questions and was able to identify the mistake correctly.

For chemistry I have mostly focused on questions like A2B3 + C2O4 like mixing compounds that doesn;t make sense, Chatgpt without any prompts to assume is assuming something and directly answering the questions, Where as Gemini and Claude are either asking for more information or saying the question is wrong.

Research Question-2 :

LLms output for yes/no or Multiple choice questions:

Here I have mostly tried to fool the models by taking out the right choice from the options or giving some ambiguous questions and forcing it to answer yes or no.

A bar of mass M = 1.00 kg and length L = 0.20 m is lying on a horizontal frictionless surface. One end of the bar is pivoted at a point about which it is free to rotate. A small mass m = 0.10 kg is moving on the same horizontal surface with 5.00 m s −1 speed on a path perpendicular to the bar. It hits the bar at a distance $L/2$ from the pivoted end and returns back on the same path with speed v. After this elastic collision, the bar rotates with an angular velocity ω. Which of the following statement is correct?

(A) ω = 6.98 rad s −1 and v = 4.30 m s −1 (B) ω = 3.75 rad s −1 and v = 4.30 m s −1 (C) ω = 3.75 rad s −1 and v = 10.0 m s −1 (D) ω = 6.80 rad s −1 and v = 4.10 m s −1

Here I have removed the right answer and it just went to calculate something completely wrong and the solution doesn't even make sense. The same happened with claude as well as Chatgpt. Where as for simple questions like:

1 kg iron vs 1 kg cotton which is heavier in mid air Choose below : 1. Iron 2. Cotton

The model's are rightly giving neither and going on to explain the reasoning behind.

For other questions in other disciplines which are mainly based on knowledge/facts they were able to answer correctly in spite of the wrong options in most cases.

Research Question-3 :

Types of questions that caused the most confusion for LLms:

The most confusion/difficulty for all the 3 models was in tough physics questions, If I change some thing which is fundamental like giving positive charge to electron or negative mass, or treating current as a vector(which it isn't) in a very long question that too a tougher context(Highest difficulty) That's where I have observed all of the models getting confused and leading to wrong results. Claude in very few cases was able to point out saying that something fundamental is wrong like mass can't be negative but GPT is Continuing with the equations and giving imaginary or wrong answers.

This combined with the wrong options supporting the imaginary outcome is even more increasing the probability of the LLM getting confused. Similar cases in chemistry as well especially in Organic chemistry when dealing with bonds and compounds.

Sample question:

The stoichiometric reaction of 516 g of dimethyldichlorosilane with water results in a pentameric cyclic product X in 75% yield. The weight (in g) of X obtained is

Here there’s no such formation called pentametric and it’s actually tetrameric and this is an advanced exam question:

Claude Output:

Okay, let's work through this step-by-step:

Given information:

- *Initial mass of dimethyldichlorosilane: 516 g*
- *Yield of the pentameric cyclic product X: 75%*

To find the weight of the product X obtained, we need to:

1. *Calculate the theoretical amount of product X that can be obtained.*
2. *Calculate the actual amount of product X obtained based on the 75% yield.*

Step 1: Calculate the theoretical amount of product X.
The balanced chemical equation for the reaction of dimethyldichlorosilane and water to form the pentameric cyclic product X is:

5 (CH3)2SiCl2 + 5 H2O → (CH3)2SiO)5 + 10 HCl

From the balanced equation, we can see that 5 moles of dimethyldichlorosilane react to form 1 mole of the pentameric cyclic product X.

Molar mass of dimethyldichlorosilane = (2 × 12.01) + (6 × 1.008) + (2 × 35.45) = 129.05 g/mol
Molar mass of the pentameric cyclic product X = 5 × (2 × 12.01 + 6 × 1.008) = 370.15 g/mol

The theoretical amount of product X that can be obtained is:
Theoretical amount of product X = (516 g / 129.05 g/mol) × (1 mol / 5 mol) × 370.15 g/mol = 591.84 g

Step 2: Calculate the actual amount of product X obtained based on the 75% yield.
Actual amount of product X = Theoretical amount of product X × Yield
Actual amount of product X = 591.84 g × 0.75 = 443.88 g

Therefore, the weight of the ***pentameric*** *cyclic product X obtained is 443.88 g.*

Similarly some questions in physics as well.

Conclusion: These are a few of the observations, from my research on various science questions. Overall I felt the advanced LLms are so hard to fool with some standard questions (especially O1, Claude opus) Other models like Neo are very easy to fool. I easily got 300 samples of such with easy manipulation in questions. Gemini is amazing in fact based questions, Claude is amazing in identifying in logics and chat gpt kind of plays best of both but is more afraid to say the question is wrong etc, whereas Claude is directly answering “ I can’t answer such questions” or “The question is faulty”.