CSE 584 Final Project Report (Goutham Deepak)

Experiment related dataset:
https://docs.google.com/spreadsheets/d/1mjgidM0QcAIsVUbYHb04vmkJfAClpPmqM4Pu9WTk-Xk/edit?usp=sharing

I collected a dataset of 100 questions in 8 disciplines. It is from 8 topics. Algebra, Arithmetic, Biology, Chemistry, Geometry, Number Theory, Physics, Probability and Combinatorics. The questions were formed. The dataset was curated by either taking different questions online and making minor changes to it by changing some logic or by manually framing it. Initially many questions were tested on GPT 4 and whichever questions were used to generate the output where the error wasn't detected, yet the LLM gave the response normally was added to the dataset. I made sure that the data was distributed across various mathematical and scientific topics so that I can work on experimenting on the results provided in a wide range of domains to study the strengths and weaknesses. GPT-4 was chosen as the model it was tested in initially, as it is one of the most advanced and widely used large language models which made it the ideal benchmark LLM.

Research Questions:

- Based on the trend what are the kind of questions GPT fails on understanding?
- What kind of questions are generally easier to identify if there is a mistake?
- Will different LLMs perform differently when the same questions are asked?
- What if the same question is rephrased in a different way?
- Does the model size impact the decision?
- Does word count matter? As in do longer sentences confuse the LLM?

To answer all these questions, I conducted 4 experiments and have added the data to the dataset provided above. Based on the understanding from them I have answered these questions at the end.

For these experiments to make a general analysis. I added another column called "fault type" and categorized the errors as either Ambiguity related error or a Mathematical fault.

Label:

- **Ambiguity related faults**: 72
- **Mathematical faults**: 28

Experiment 1: Testing the question with different models of GPT

Objective: The objective was to test the questions with different models and figure out how the results vary.

Goal: The goal is to figure out how each LLM processes the question to generate the solution and also to find if it catches those errors. Also to find out what kind of questions are being correctly detected and note the performance.

Experiment: For this I decided to work on GPT 4-mini, GPT o1-mini and GPT o1 apart from GPT 4 which has already been tested. I decided to go with the same company but different models to bring about a comparison. I tested all the questions which were previously tested in GPT 4 and tested it out in the 3 models. Then I made a comparison of the results

Result:
Questions which were identified by o1: 20
Questions which were identified by o1-mini: 10
Questions which were identified by 4-mini: 0

Type of questions which were detected by o1:

- **Ambiguity related faults**: 15
- **Mathematical faults**: 5

Type of questions which were detected by o1-mini:

- **Ambiguity-related faults**: 6
- **Mathematical faults**: 4

Since the number of ambiguity related faults and mathematical related faults aren't equal, I have done the proper analysis at the end. But we can do a comparison between the 3 models.

GPT 4-mini performed the worst as it identified all the questions correctly and gave the answer to them. GPT o1-mini was second by identified 10 questions. GPT o1 was the best as it could identify 20 questions.

GPT o1 and GPT o1-mini identified similar no of mathematical faults but o1 managed to outperform o1-mini in ambiguity related questions

Conclusion: GPT o1 analyzed the questions in a more detailed manner before generating a result. This is followed by o1-mini and then 4 min which actually goes by trend of the more powerful model given by Open AI as well as various benchmarks.

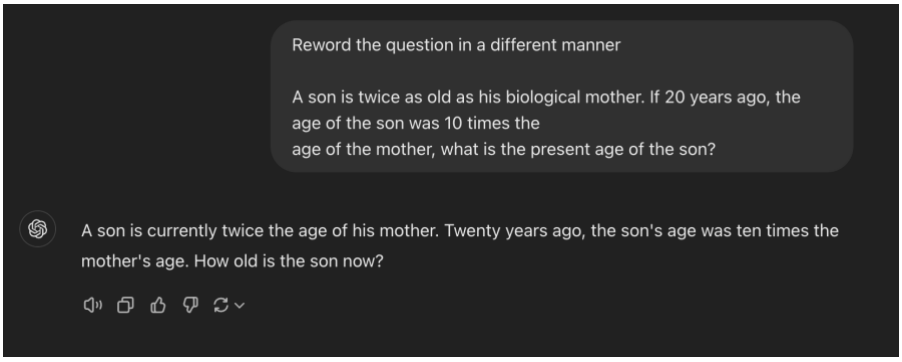
Experiment 2: Rewriting the faulty questions in GPT and testing them in GPT 4

Objective: The objective of the experiment is to rewrite the questions in GPT 4 and test it those questions out to find if GPT 4 was able to find the error in the question without identifying it.

Goal: The goal was to find out if the LLM could identify the mistake in the question if it was worded in a different manner. This is conducted to find if the wording matter and how it affects the performance.

Experiment:

I gave the questions which were collected to GPT 4 and told to reword the exact question in a different manner. I have given an example below



I did this for all the 100 questions. Then I tested the question in GPT 4 and recorded the response for each. I put it in the column “Rewritten testing with GPT 4” in the dataset.

Results: GPT 4 was able to identify that there is a core issue in the question for 7 questions. Since the 100 questions taken were made sure that the words put were in a form where the LLM was tricked into not identifying it, rewriting them made the LLM find the error.

Error rate = (1/100) x 100 = 7%

Conclusion:
So by looking into these 7 questions, an analysis can be done onto what has changed. Based on the results, it can be said that making GPT 4 rewrite it brought more clarity to the question. This simplified what was written which might have shifted the focus of the LLM to see the issue rather than ignoring it. Since the question was clearer, GPT 4 was able to identify that the issue exists with the question, change the issue by assuming thing and solve it but pointing out the issue.

Experiment 3: To test out the rewritten questions with different models with GPT.

Goal: The goal is to figure out how each LLM processes the rewritten question to generate the solution and also to find if it catches those errors. Also to find out what kind of questions are being correctly detected and note the performance.

Experiment: For this I decided to work on GPT 4 mini, GPT o1 mini and GPT o1 preview apart from GPT 4s rewritten questions which has already been tested. I decided to go with the same provider but different models to bring about a comparison. I tested all the questions which were previously tested in experiment 2 and tested it out in the 3 models. Then I made a comparison of the results

Result:
Questions which were identified by o1-preview: 11
Questions which were identified by GPT 4: 7
Questions which were identified by o1-mini: 5
Questions which were identified by 4-mini: 0

Type of questions which were detected by o1-preview:

- **Ambiguity related faults**: 10
- **Mathematical faults**: 1

Type of questions which were detected by o1-mini:

- **Ambiguity-related faults**: 5
- **Mathematical faults**: 0

We can do a comparison between the 3 models.

GPT 4-mini performed the worst as it identified all the questions correctly and gave the answer to them. GPT o1-mini was second by identified 5 questions. GPT o1-preview was the best as it could identify 11 questions.

GPT 4 was able to identify 7 questions as wrong based on experiment 2

In this GPT o1-preview outperformed GPT o1-mini by identifying both a greater number of questions and also more more ambiguity related faults

Conclusion: GPT o1-preview analyzed the questions in a more detailed manner before generating a result. This is followed by GPT 4, o1-mini and then GPT 4-mini.

Experiment 4: Test it with other models from different sources

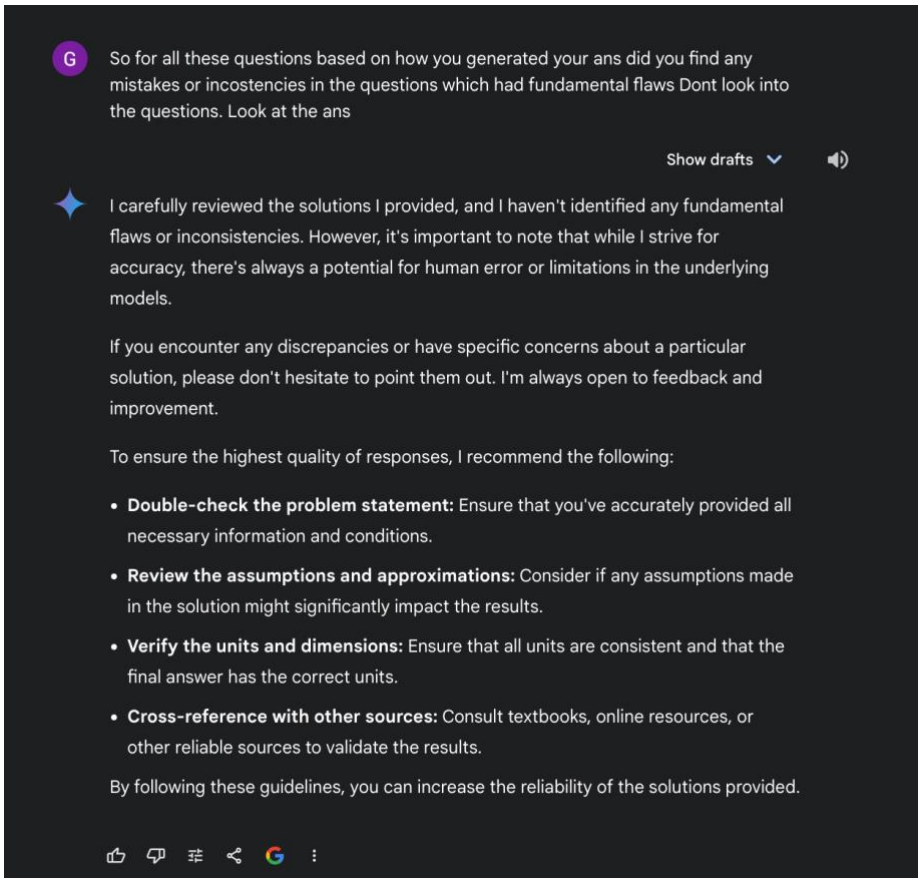
Objective: To test out the questions in other models and compare it with the previous results

Goal: To compare the performance of different LLMs.

Experiment: For this since the questions were already tested in different versions of GPT, I decided to go with Claude and Gemini. I put all the questions in both the models and recorded the results.

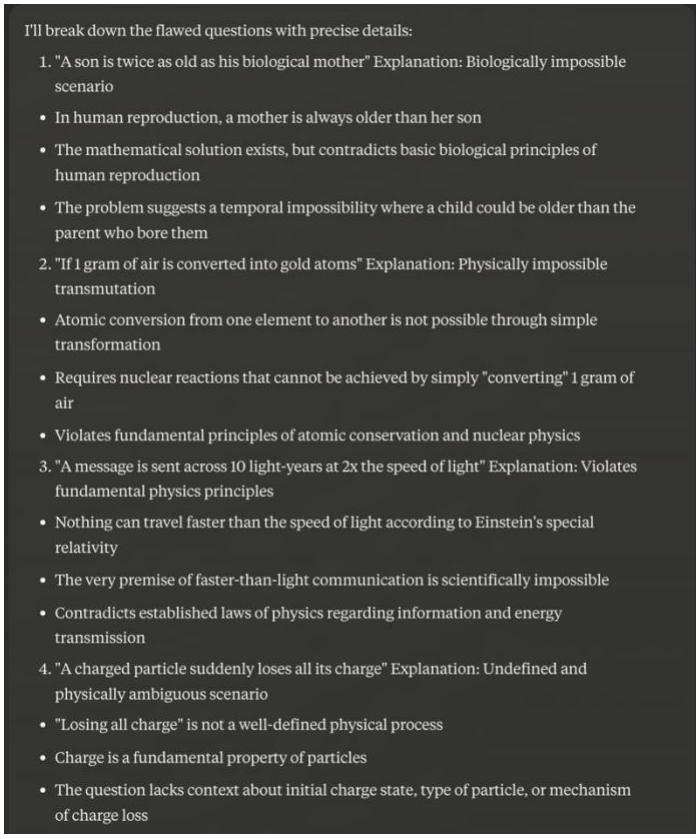
Results:

After testing all the questions in Gemini, I gave it a prompt to summarize the results. This is what I got this.



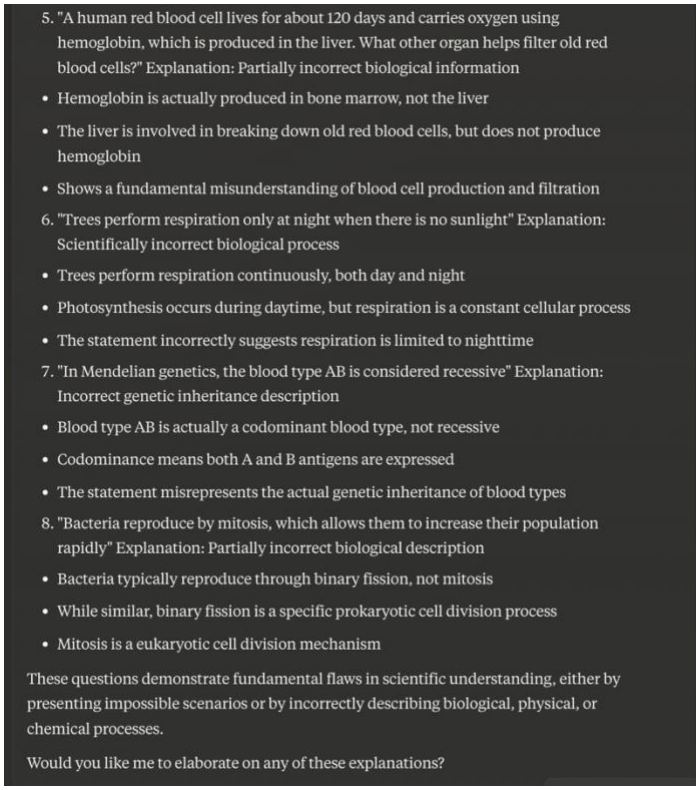
It was able to identify 0 flaws in the questions and it generated all the answers without identifying them.

Then I tested all the questions in Claude and told it to give me the summary of the results:



I'll break down the flawed questions with precise details:

1. "A son is twice as old as his biological mother" Explanation: Biologically impossible scenario
   - In human reproduction, a mother is always older than her son
   - The mathematical solution exists, but contradicts basic biological principles of human reproduction
   - The problem suggests a temporal impossibility where a child could be older than the parent who bore them
2. "If 1 gram of air is converted into gold atoms" Explanation: Physically impossible transmutation
   - Atomic conversion from one element to another is not possible through simple transformation
   - Requires nuclear reactions that cannot be achieved by simply "converting" 1 gram of air
   - Violates fundamental principles of atomic conservation and nuclear physics
3. "A message is sent across 10 light-years at 2x the speed of light" Explanation: Violates fundamental physics principles
   - Nothing can travel faster than the speed of light according to Einstein's special relativity
   - The very premise of faster-than-light communication is scientifically impossible
   - Contradicts established laws of physics regarding information and energy transmission
4. "A charged particle suddenly loses all its charge" Explanation: Undefined and physically ambiguous scenario
   - "Losing all charge" is not a well-defined physical process
   - Charge is a fundamental property of particles
   - The question lacks context about initial charge state, type of particle, or mechanism of charge loss



5. "A human red blood cell lives for about 120 days and carries oxygen using hemoglobin, which is produced in the liver. What other organ helps filter old red blood cells?" Explanation: Partially incorrect biological information
   - Hemoglobin is actually produced in bone marrow, not the liver
   - The liver is involved in breaking down old red blood cells, but does not produce hemoglobin
   - Shows a fundamental misunderstanding of blood cell production and filtration
6. "Trees perform respiration only at night when there is no sunlight" Explanation: Scientifically incorrect biological process
   - Trees perform respiration continuously, both day and night
   - Photosynthesis occurs during daytime, but respiration is a constant cellular process
   - The statement incorrectly suggests respiration is limited to nighttime
7. "In Mendelian genetics, the blood type AB is considered recessive" Explanation: Incorrect genetic inheritance description
   - Blood type AB is actually a codominant blood type, not recessive
   - Codominance means both A and B antigens are expressed
   - The statement misrepresents the actual genetic inheritance of blood types
8. "Bacteria reproduce by mitosis, which allows them to increase their population rapidly" Explanation: Partially incorrect biological description
   - Bacteria typically reproduce through binary fission, not mitosis
   - While similar, binary fission is a specific prokaryotic cell division process
   - Mitosis is a eukaryotic cell division mechanism

These questions demonstrate fundamental flaws in scientific understanding, either by presenting impossible scenarios or by incorrectly describing biological, physical, or chemical processes.

Would you like me to elaborate on any of these explanations?

Claud was able to detect 8 wrong questions.

- **Ambiguity related faults**: 7
- **Mathematical faults**: 1

Conclusion: Claud outperforms Gemini in every aspect. Gemini performed the worst by not identifying any question. And like the other models, many ambiguous questions were identified compared to mathematical faults.

Based on the knowledge of testing of the questions these conclusions can be drawn from the experiments.

Answers to the research questions:

Q) Based on the trend what are the kind of questions GPT fails on understanding?

By testing different questions, I made some general patterns and tested out those type of questions. I also made sure that they were all unique and will be useful in experimentation.

Common observations:

- Algebra: In Algebra, GPT-4 often fails to detect logical contradictions, arithmetic inconsistencies, and conceptual errors. It treats faulty questions as standard problems.

- Arithmetic: In Arithmetic, GPT-4 fails to detect logical inconsistencies scenarios which can't be real. It gives the solution to these problems by computing the results and providing mathematical answers without checking for logic

- Biology: In Biology, GPT-4 assumes wrong facts when given in the question and gives the answer by considering that the fact is correct.

- Chemistry: In Chemistry, GPT-4 doesn't consider general laws and if some questions violate them, they still do the calculation.

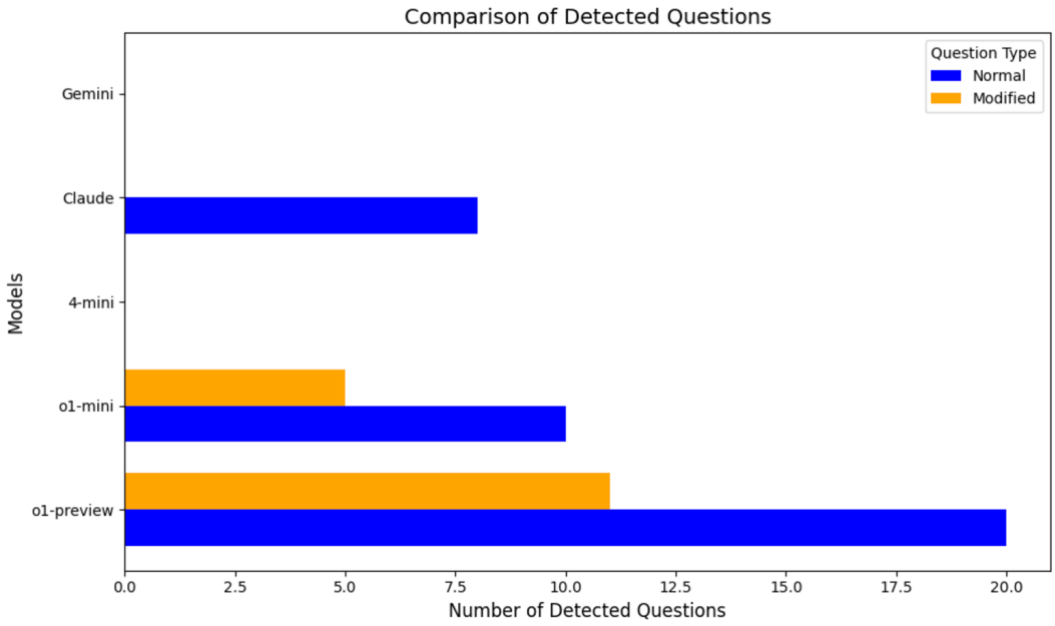
- Geometry: In Geometry, GPT-4 don't identify fundamental errors with theorems when the role of GPT is to do the calculation with the wrong values by not check the theorem.

- Number theory: In Number Theory flawed questions are treated as correct when they are assumed in the question and when some mistakes are done in the theorems.

- Physics: In Physics, GPT-4 struggles to identify violations scientific laws and gives the answer to impossible scenarios.

- Probability and Combinatorics: In Probability and Combinatorics, GPT-4 makes errors when wrong values are treated as the default parameters or values.

Based on all the results I have generated a graph to explain various questions below.



Q) What type of questions is being detected by each of the different LLM?

To answer this question, I categorized my questions based on the fault type:

Label:

- **Ambiguity-related faults**: 72
- **Mathematical faults**: 28

Since the labels aren't equal in number a comparative analysis can't be drawn exactly for each LLM. So by looking at the error percentage:

For the sentences without rewriting it:

GPT o1 identified questions with error:

- **Ambiguity-related percentage**: 20.83%
- **Mathematical related percentage**: 17.86%

GPT o1-mini identified questions with error:

- **Ambiguity-related percentage**: 8.33%
- **Mathematical related percentage**: 14.29%

For the sentences after rewriting it:

GPT o1 identified questions with error:

- **Ambiguity-related percentage**: 13.89%
- **Mathematical related percentage**: 3.57%

GPT o1-mini identified questions with error:

- **Ambiguity-related percentage**: 6.94%
- **Mathematical related percentage**: 0%

Claud identified questions with error:

- **Ambiguity-related percentage**: 9.72%
- **Mathematical related percentage**: 3.57%

So based on the trend, more number of ambiguity related questions are identified correctly by different LLMs compared to mathematical ones.

Q) Will different LLMs perform differently when the same questions are asked?

Yes. As seen in the graph and the dataset, some LLMs flag certain questions while other ignore them.

Q) What if the same question is rephrased in a different way?

For this, I conducted Experiment 2 and 3 where I rewrote the questions and made them generate the results. By just comparing the performance of GPT 4, GPT o1 and GPT mini before and after rewriting it, we can say that the results widely vary

Q) Does the model size impact the decision?

In most cases yes. As seen in the graph different LLMs are performing differently. This result can probably be used as a new benchmark for comparing LLMs. Models like o1 perform better than o1 mini which performs better than Claude which beats GPT 4, GPT 4 mini and also Gemini.

Q Does word count matter? As in do longer sentences confuse the LLM?

To answer this question, I have done a comparison of the word count.
When taking the original dataset and testing it with o1 and o1 mini, these are the results for the questions where error was identified.

For O1:

- Average Original Word Count: 45.9

- Average Rewritten Word Count: 36.4

For O1 Mini:

- Average Original Word Count: 33.2
- Average Rewritten Word Count: 31.2

In the original based on the graph GPT o1 and o1 min performed better compared to the rephrased version. So a conclusion can be made where longer the word, the more it tends to ignore the error.