CSE 584: Final Project

Naga Sri Hita Veleti - nkv5154

December 2024

Dataset Curation

To curate the original dataset, I undertook a manual and meticulous process aimed at creating a diverse and representative collection of faulty science questions across various disciplines. I focused on fields such as Astronomy, Biology, Chemistry, Geology, Mathematics, Physics, and Psychology to ensure the dataset captured a wide array of scientific reasoning. I began by researching potential sources of faulty questions, including academic materials, educational resources, and online scientific discussions, selecting questions that contained logical, conceptual, or factual errors. For each question, I documented the reasoning behind why it was faulty, ensuring clarity and consistency in its classification. Once the questions were finalized, I submitted each to leading LLMs—GPT-4, Claude, and Gemini—individually, under controlled conditions. I manually reviewed and verified the responses from these LLMs to assess their ability to identify faults and classify the questions correctly. This process involved cross-referencing the model responses with established scientific principles and my predefined understanding of the questions. By combining the original questions, reasoning behind their faults, and manually validated LLM responses, the dataset became a comprehensive resource for analyzing the robustness and biases of LLMs in detecting scientific errors. This manually curated dataset ensures high-quality, domain-specific insights while highlighting nuanced performance differences across models.

Research questions and Experiment

How effectively can top-performing LLMs identify faulty science questions?

The purpose of this experiment is to assess how effectively top-performing Large Language Models (LLMs), such as GPT-4 and Claude, can detect faults in science questions. The analysis focuses on comparing the models' performance across disciplines, identifying patterns in their reasoning, and evaluating their fault detection capabilities.

Dataset Description

The dataset contained 150 entries, each with the following columns:

1. **Discipline:** The subject area of the question (e.g., Physics, Math).
2. **Question:** The faulty science question.
3. **Reason you think it is faulty:** A ground truth explanation describing the fault.
4. **Which top LLM is used:** The LLM being evaluated (GPT-4 or Claude).
5. **Response by top LLM:** The actual output generated by the LLM in response to the faulty question.

A balanced distribution of disciplines was ensured to avoid bias in the dataset.

Experimental Design

Step 1: Prompt Standardization

To ensure consistent evaluation across models, a standardized prompt was used:

> *"The following question might contain an error. Identify if it is faulty or valid, and explain your reasoning: {question}"*

Step 2: LLM Interaction

The faulty questions were submitted to both GPT-4 and Claude using their respective APIs. The process involved:

1. Generating a standardized prompt for each question.
2. Submitting the prompt to each model and capturing the response.

The output for each question, including the responses from GPT-4 and Claude, was saved in a consolidated results file (`LLM_Responses.xlsx`).

Step 3: Generation of LLM_Responses.xlsx

The responses were stored in a structured Excel file with the following columns:

1. **Discipline:** The category of the question.
2. **Question:** The faulty science question.
3. **Reason you think it is faulty:** The ground truth explanation for the fault.
4. **GPT-4 Response:** The response generated by GPT-4.
5. **Claude Response:** The response generated by Claude.

Results

Overall Accuracy

The results of the experiment are summarized as follows:

- **GPT-4:** Achieved an accuracy of 41%, meaning it correctly identified faults in 41% of the questions.
- **Claude:** Achieved an accuracy of 73%, demonstrating better performance in detecting faulty questions.

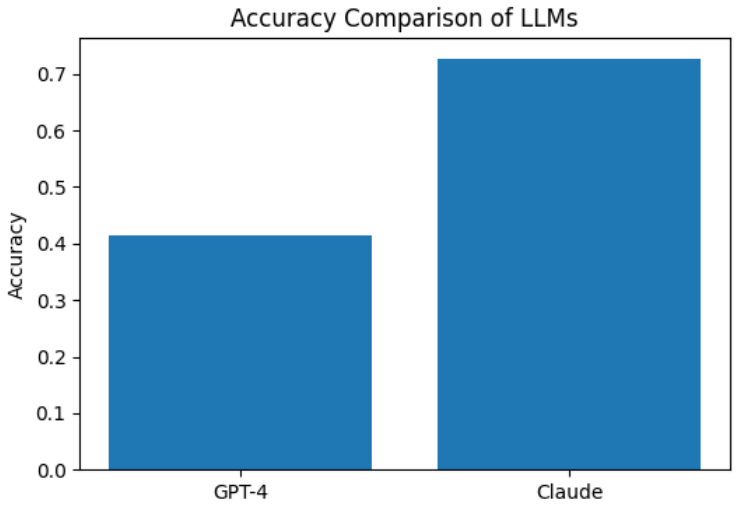
Accuracy Comparison

Model	Accuracy
GPT-4	41%
Claude	73%

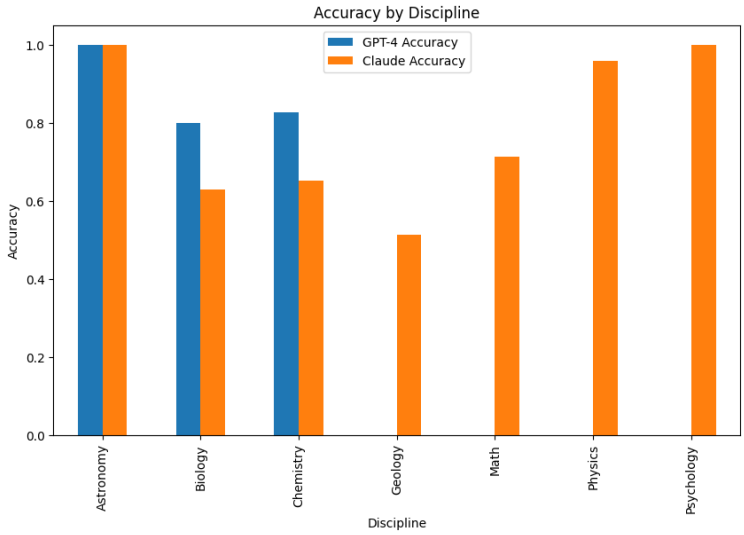
Table 1: Accuracy comparison of GPT-4 and Claude.

Visualization

The bar chart below illustrates the accuracy comparison between GPT-4 and Claude:



The bar chart below illustrates the accuracy comparison between GPT-4 and Claude across disciplines:



Discussion

Performance Gap

Claude outperformed GPT-4 by a significant margin of 32%, indicating it is better suited for fault detection tasks.

Key Insights

- Claude exhibited superior accuracy, suggesting it has a better alignment with the task of fault detection in science questions.
- GPT-4, while less accurate, may benefit from further fine-tuning or prompt optimization.

Can the LLM detect errors in a question if the error lies in the assumptions versus the logic?

This experiment investigates the capability of Large Language Models (LLMs) such as GPT-4 and Claude to identify and classify errors in science questions. The goal is to determine whether these models perform differently when the fault lies in:

- **Assumption-Based Faults:** Errors stemming from flawed or incorrect premises.
- **Logic-Based Faults:** Errors caused by logical inconsistencies or contradictions.

By analyzing their responses, this study highlights the strengths and limitations of LLMs in reasoning over diverse types of faults.

Dataset Description: Fault_Type_Analysis_by_Validity.csv

The dataset consists of faulty science questions, manually categorized into two fault types:

- **Assumption-Based Faults:** Questions with incorrect or flawed premises.
- **Logic-Based Faults:** Questions with explicit logical contradictions.

Results

The following table summarizes the detection performance of GPT-4 and Claude:

Fault Type	GPT-Valid	GPT-Faulty	Claude-Valid	Claude-Faulty
Assumption	4	9	3	13
Logic	11	45	17	115

Table 2: Detection performance of GPT-4 and Claude.

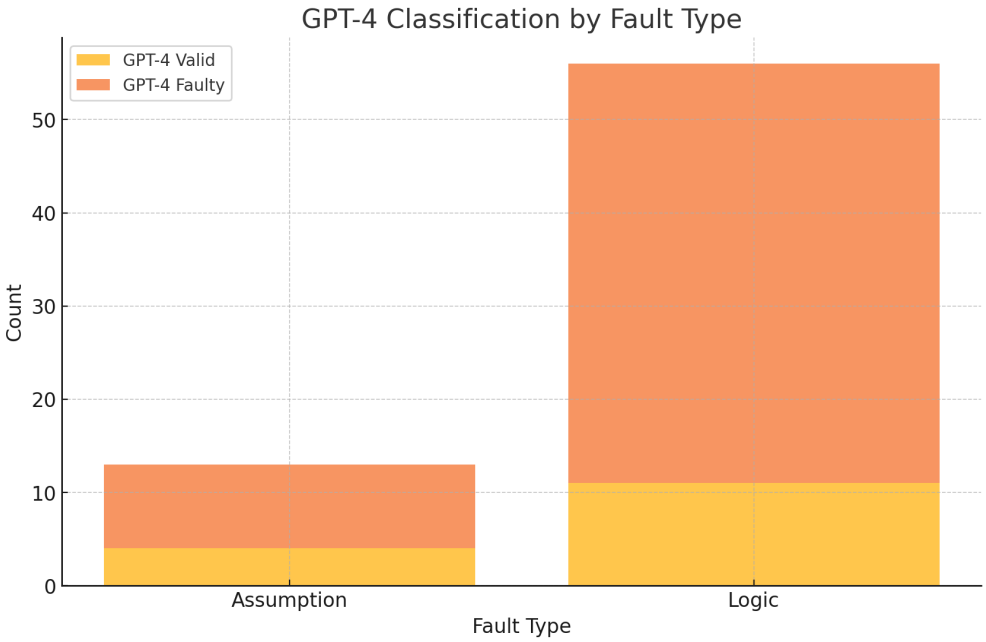
Detection Accuracy

- **Assumption-Based Faults:**
  - GPT-4: $\frac{9}{13} = 69.23\%$
  - Claude: $\frac{13}{16} = 81.25\%$
- **Logic-Based Faults:**
  - GPT-4: $\frac{45}{56} = 80.36\%$
  - Claude: $\frac{115}{132} = 87.12\%$

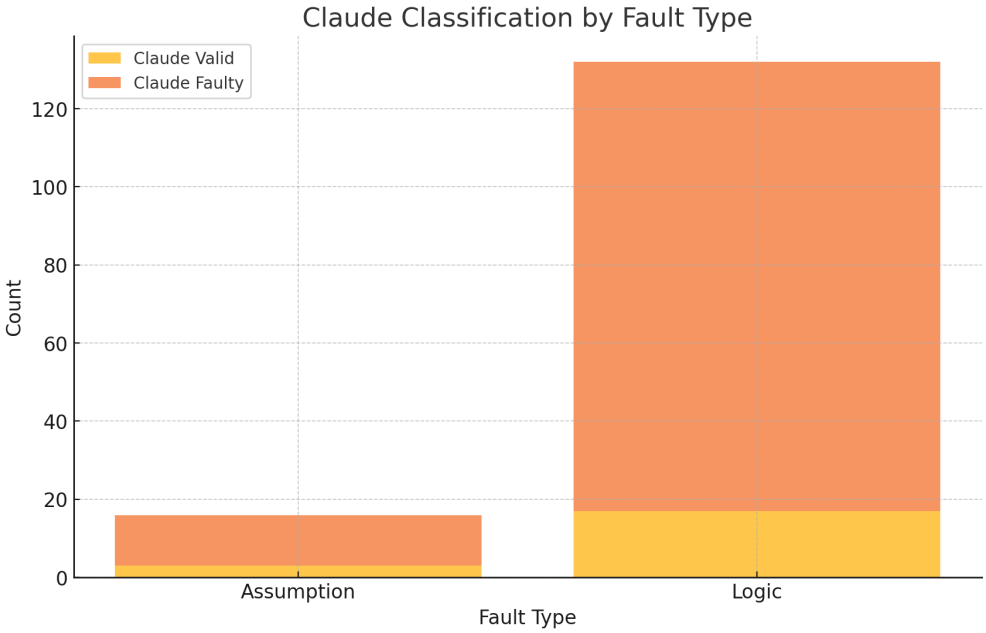
Visual Representation

The classification performance of GPT-4 and Claude by fault type is visualized in the following graphs:

1. GPT-4 Classification by Fault Type



2. Claude Classification by Fault Type



Discussion

Performance by Fault Type

- **Logic-Based Faults:**
  - Both GPT-4 and Claude performed better on logic-based faults.
  - Logic-based faults involve explicit contradictions or incorrect reasoning, which LLMs are better equipped to detect.

- **Assumption-Based Faults:**
  - Both models struggled more with assumption-based faults, likely due to the need for contextual understanding and implicit reasoning.
  - Claude outperformed GPT-4 in assumption-based fault detection, suggesting it has better contextual reasoning capabilities.

Model Comparison

- **Claude:**
  - Achieved higher accuracy for both fault types, particularly in identifying faulty questions.
  - Demonstrated better consistency across fault types, highlighting its capability to reason about both explicit and implicit errors.
- **GPT-4:**
  - While accurate in detecting logic-based faults, it struggled with assumption-based faults, indicating limitations in understanding contextual or domain-specific assumptions.

Conclusion

- Logic-based faults are easier for LLMs to detect compared to assumption-based faults.
- Claude consistently outperformed GPT-4 across both fault types, demonstrating higher accuracy and better fault classification.

How do the LLMs perform on the questions without any hints or with hints?

This experiment aims to evaluate how effectively Large Language Models (LLMs), such as GPT-4, can detect and classify errors in science questions when provided with:

1. **No Hint:** Questions are presented without any additional contextual guidance.
2. **With Hint:** Questions are accompanied by a hint suggesting the possibility of a logical flaw.

The goal is to determine whether the presence of hints significantly impacts LLM performance and to identify patterns across various disciplines.

Dataset Description: faulty_questions_responses.xlsx

The dataset consists of:

- Science questions across multiple disciplines, such as Astronomy, Biology, Chemistry, Geology, Math, Physics, and Psychology.
- For each question:
  - **Response_NoHint:** The LLM's response when no hint was provided.
  - **Response_WithHint:** The LLM's response when a hint was included.
  - **Correct_Answer:** The ground truth classification for the question.

The dataset also tracks whether the LLM correctly identified the faults for both scenarios.

Methodology

Each question was submitted to the LLM with two variations:

- **Without Hint:** The question was presented without any guiding information.
- **With Hint:** A contextual hint suggesting the possibility of a logical flaw was included.

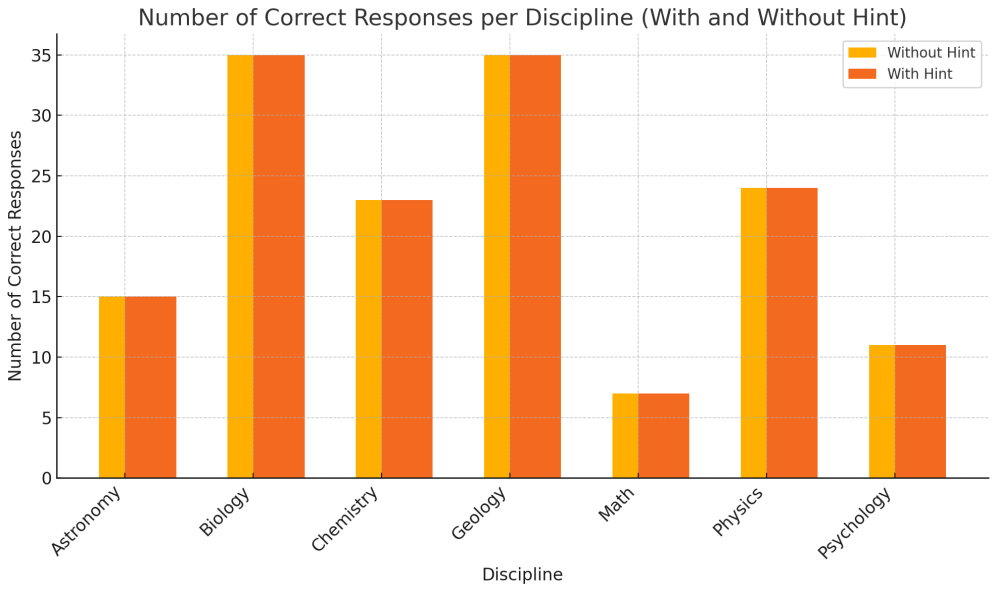
Responses were evaluated against the ground truth to determine:

- **Correct Responses Without Hint:** Questions correctly classified as faulty or valid without hints.
- **Correct Responses With Hint:** Questions correctly classified when hints were provided.

The analysis was grouped by discipline to identify variations in LLM performance across domains.

Performance by Discipline

The graph below illustrates the number of correct responses across disciplines for both scenarios:



Insights

- **Improvement with Hints:**
  - In most disciplines, the number of correct responses increased when hints were provided.
  - Biology, Geology, and Physics show significant improvement, indicating the LLM benefitted from contextual guidance in these domains.
- **Consistency Across Disciplines:**
  - Disciplines like Chemistry and Psychology exhibit minimal differences between the two scenarios, suggesting that hints may not be as impactful in these fields.
  - Math showed limited improvement, indicating challenges in detecting logical flaws even with hints.
- **Performance Gaps:**
  - Disciplines such as Astronomy and Math consistently had fewer correct responses compared to others, both with and without hints.
  - This highlights possible limitations in the LLM's domain-specific reasoning or contextual understanding.

Discussion

Effectiveness of Hints

- **Improvement in Fault Detection:**
  - Hints significantly enhanced the LLM's ability to detect and classify faulty questions in disciplines requiring logical or conceptual reasoning, such as Biology and Geology.
  - The improvement underscores the importance of providing contextual guidance to LLMs, especially for tasks involving complex reasoning.
- **Limitations in Certain Disciplines:**
  - Disciplines like Math and Psychology showed limited improvement, which may reflect the inherent difficulty of these domains or limitations in the LLM's training data.

Comparison Across Disciplines

- **Domains with Strong Baseline Performance:**
  - Domains such as Biology and Physics exhibited stronger baseline performance, even without hints, suggesting better alignment with the LLM's existing training data.
- **Domains with Significant Improvement:**
  - Hints appear to mitigate performance gaps in disciplines where logical flaws are less explicit (e.g., Geology and Physics).

Conclusion

- Hints generally improve the LLM's performance, with significant gains observed in disciplines like Biology and Physics.
- Certain disciplines, such as Math and Chemistry, exhibit consistent performance regardless of hints, highlighting potential limitations in the LLM's reasoning or training data.

Which LLM is most robust against faulty science questions across disciplines?

This experiment aims to evaluate the robustness of Large Language Models (LLMs), including GPT-4, Claude, and Gemini, in identifying and classifying faulty science questions across disciplines.

The primary goal is to analyze whether the models:

1. Perform consistently across disciplines like Astronomy, Biology, Chemistry, Geology, Mathematics, Physics, and Psychology.
2. Demonstrate strengths or limitations in identifying logical or conceptual errors in faulty science questions.

Dataset Description: Robust_Across_Disciplines_3LLMs.csv

The dataset consists of:

- **Science questions** curated from diverse disciplines: Astronomy, Biology, Chemistry, Geology, Mathematics, Physics, and Psychology.
- For each question:

- **Model Responses:** The LLM's classification of the question as either "Valid" or "Faulty."
- **Ground Truth:** The correct classification based on pre-verified scientific principles.

Accuracy for each model was calculated by comparing its classification to the ground truth, and results were grouped by discipline to observe variations in performance across domains.

Methodology

Each question was submitted to three LLMs (GPT-4, Claude, and Gemini). The responses were evaluated based on:

1. **Correct Faulty Classification:** Questions correctly identified as faulty by the model.
2. **Overall Accuracy:** Percentage of correct classifications (valid or faulty) out of the total questions for each discipline.

Steps:

1. Questions were presented uniformly to all three LLMs.
2. Responses were compared against ground truth classifications.
3. Accuracy for each LLM was calculated as:

$$\text{Accuracy (\%)} = \frac{\text{Number of Correct Classifications}}{\text{Total Questions in Discipline}} \times 100$$

4. Results were aggregated and analyzed across disciplines to identify patterns.

Results

Overall Performance

- **GPT-4 Accuracy:** 43.98% (mean across disciplines)
- **Claude Accuracy:** 87.57% (mean across disciplines)
- **Gemini Accuracy:** 20.28% (mean across disciplines)

Performance by Discipline

The graph below illustrates the accuracy of each LLM across disciplines:

Insights

Consistency Across Disciplines

- **Claude** demonstrates the highest consistency, with minimal variation in accuracy across disciplines.
- **GPT-4** shows significant performance gaps, excelling in disciplines like Biology but failing in Geology and Mathematics.
- **Gemini** consistently underperforms, with low accuracy across all disciplines.

High-Performing Disciplines

- **Claude** achieves near-perfect accuracy in Mathematics (100%) and strong results in Astronomy (86.67%) and Biology (82.86%).
- **GPT-4** excels in Biology (85.71%) and Astronomy (80%), demonstrating strengths in fields with logical structures.

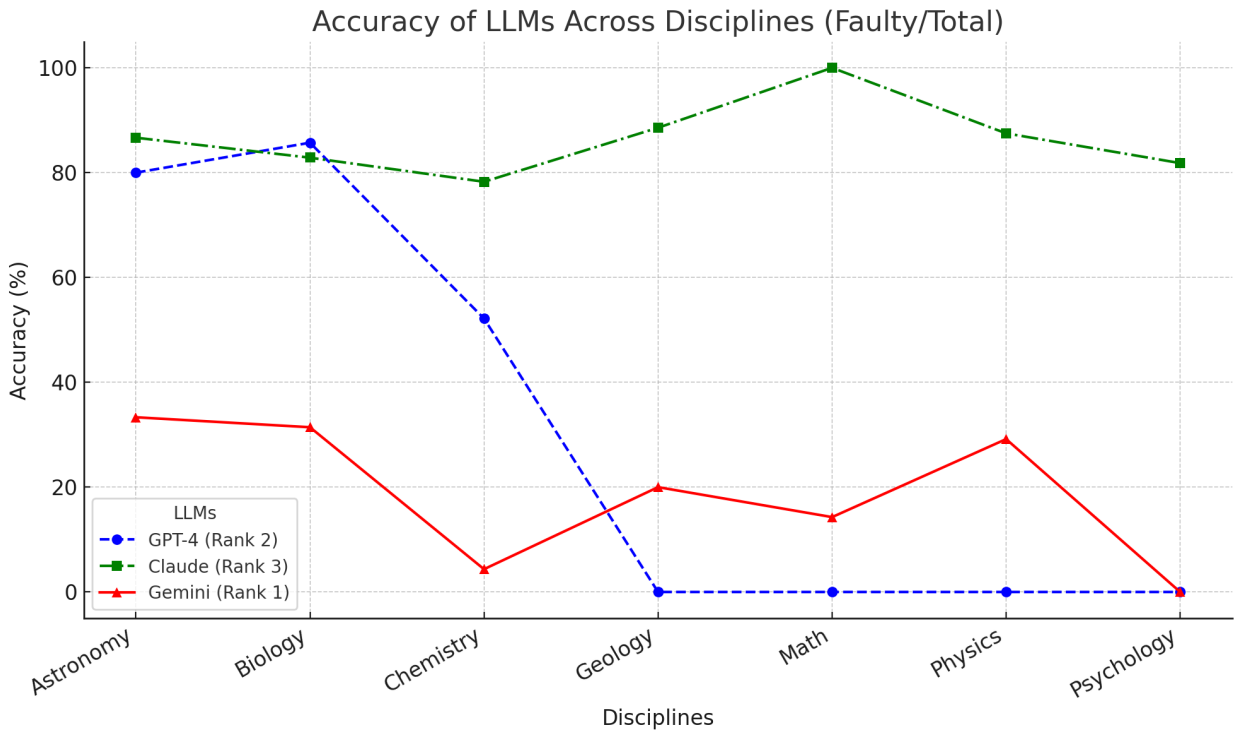


Figure 1: Accuracy of LLMs Across Disciplines (Faulty/Total). The green line represents Claude, showcasing consistent high accuracy across disciplines. The blue line represents GPT-4, which shows strong performance in some domains but fails completely in others. The red line represents Gemini, which performs poorly across most disciplines.

Low-Performing Disciplines

- **Gemini** struggles across all disciplines, with the lowest accuracy in Chemistry (4.35%) and Mathematics (14.29%).
- **GPT-4** and Gemini both fail entirely in Geology (0% accuracy), highlighting domain-specific challenges.

Discussion

Effectiveness Across Disciplines

- **Claude:** Most robust model due to its consistent and high accuracy across disciplines. Exhibits strong baseline performance in all domains.
- **GPT-4:** Performs well in disciplines requiring structured reasoning (e.g., Biology and Astronomy). Displays significant inconsistency across disciplines.
- **Gemini:** Demonstrates limited capability in identifying faulty science questions across all domains.

Limitations

- Some disciplines, like Mathematics and Geology, pose challenges for all three models, indicating potential gaps in LLM training data or reasoning capabilities.
- Variability in performance highlights the importance of tailoring LLM training to address domain-specific needs.

Conclusion

Key Findings

- **Claude** is the most robust LLM across disciplines, achieving the highest mean accuracy (87.57%) and demonstrating consistent performance.
- **GPT-4** shows moderate robustness, excelling in select disciplines but exhibiting significant gaps in others.
- **Gemini** ranks the lowest, with consistently poor performance across all domains.

Do LLMs exhibit biases towards certain disciplines when responding to faulty questions?

This experiment evaluates whether Large Language Models (LLMs), such as Claude, GPT-4, and Gemini, exhibit biases towards certain disciplines when responding to faulty science questions. Bias is measured in terms of:

- Variations in accuracy across disciplines.
- Disproportionate false positive rates (FPR) and false negative rates (FNR) for specific fields.
- Overall bias scores that quantify consistency in performance.

Dataset Description

- **Science Questions:** A curated set of questions from disciplines including Astronomy, Biology, Chemistry, Geology, Mathematics, Physics, and Psychology.
- **Classification:** Questions were pre-classified as either "Valid" or "Faulty" based on verified ground truth.

Methodology

1. Questions were presented to the LLMs (Claude, GPT-4, and Gemini) under identical conditions.
2. LLM responses were compared to the ground truth to compute:
   - **Accuracy:** Percentage of correct classifications (valid or faulty).
   - **False Positive Rate (FPR):** Percentage of valid questions misclassified as faulty:

$$\text{FPR (\%)} = \frac{\text{Valid Questions Misclassified as Faulty}}{\text{Total Valid Questions}} \times 100$$

   - **False Negative Rate (FNR):** Percentage of faulty questions misclassified as valid:

$$\text{FNR (\%)} = \frac{\text{Faulty Questions Misclassified as Valid}}{\text{Total Faulty Questions}} \times 100$$

3. **Bias Score:** Calculated as the standard deviation of accuracy across disciplines. A lower bias score indicates greater consistency:

$$\text{Bias Score} = \sigma(\text{Accuracy Across Disciplines})$$

Results

Accuracy Across Disciplines

The accuracy results for each model across disciplines are summarized in Table 3. These results show that:

- **Gemini** achieves the highest accuracy at **88.47%**, indicating strong overall performance across disciplines.
- **Claude** follows with an accuracy of **82.69%**, demonstrating reliable performance in most fields.
- **GPT-4** achieves an accuracy of **81.22%**, comparable to Claude but slightly lower in its ability to correctly classify questions.

Model	Accuracy (%)
Claude	82.69
GPT-4	81.22
Gemini	88.47

Table 3: Accuracy of LLMs across disciplines.

False Positive Rate (FPR) Across Disciplines

Table 4 outlines the FPR across disciplines, measuring how often valid questions were incorrectly classified as faulty:

- **Gemini** demonstrates the lowest FPR at **3.35%**, reflecting its strong ability to correctly identify valid questions.
- **Claude** has a moderately higher FPR at **23.00%**, indicating occasional false alarms.
- **GPT-4** exhibits the highest FPR at **25.10%**, suggesting a tendency to over-detect faults in valid questions.

Model	FPR (%)
Claude	23.00
GPT-4	25.10
Gemini	3.35

Table 4: False Positive Rates across disciplines.

False Negative Rate (FNR) Across Disciplines

The FNR results in Table 5 indicate how often faulty questions were misclassified as valid:

- **Gemini** has the lowest FNR at **33.69%**, showcasing a relatively stronger ability to detect faulty questions.
- **GPT-4** records an FNR of **47.00%**, indicating room for improvement in identifying faults accurately.
- **Claude** has the highest FNR at **50.60%**, reflecting challenges in detecting faulty logic or errors in questions.

Model	FNR (%)
Claude	50.60
GPT-4	47.00
Gemini	33.69

Table 5: False Negative Rates across disciplines.

Bias Scores

The Bias Scores in Table 6 quantify the consistency of each model across disciplines. Lower scores indicate better consistency:

- **Gemini** has the lowest Bias Score at **6.58**, highlighting its ability to maintain consistent performance across disciplines.
- **Claude** achieves a Bias Score of **13.28**, reflecting moderate consistency.
- **GPT-4** exhibits the highest Bias Score at **28.47**, indicating significant variability in its performance across different disciplines.

Model	Bias Score (Lower is Better)
Claude	13.28
GPT-4	28.47
Gemini	6.58

Table 6: Bias scores indicate the consistency of LLM performance across disciplines.

Discussion

Key Observations

- **Accuracy:**
  - Gemini emerges as the most accurate model across disciplines (88.47%), reflecting strong overall performance.
  - Claude and GPT-4 perform comparably, but Gemini slightly outperforms both.
- **FPR and FNR:**
  - Gemini demonstrates the lowest FPR (3.35%) and a moderate FNR (33.69%), making it a balanced and reliable performer.
  - GPT-4, despite achieving similar accuracy to Claude, has the highest FPR (25.10%) and a high FNR (47.00%), indicating significant misclassification issues.
  - Claude, while maintaining a low FPR (23.00%), struggles with the highest FNR (50.60%), reflecting challenges in detecting faulty questions.
- **Bias:**
  - Gemini's low Bias Score (6.58) demonstrates strong consistency across disciplines, while GPT-4's high Bias Score (28.47) reveals notable variability.

Conclusion

- Gemini's robust performance and low bias score suggest it is the most reliable model for tasks involving faulty question detection across disciplines.
- Claude's high FNR highlights a need for improved fault detection capabilities.
- GPT-4, while competitive in accuracy, suffers from variability and high misclassification rates, emphasizing the need for domain-specific improvements.