Project Report: Evaluating Large Language Models' Performance on Faulty and Misleading Questions Across Disciplines

Name : Raghavendra Jagirdar
PSU ID : 911938232

Introduction

Large Language Models (LLMs) like GPT-4, ChatGPT, Gemini-1.5-Pro, and Claude-3-Opus have revolutionized the way we interact with artificial intelligence, offering capabilities ranging from natural language understanding to complex problem-solving. However, the effectiveness of these models can be challenged by faulty or misleading questions that may confuse their algorithms, leading to incorrect or nonsensical responses. This project aims to evaluate the performance of top-tier LLMs when confronted with such challenging questions across various academic disciplines, including Astronomy, Mathematics, Physics, Chemistry, Biology, Earth Science, Psychology, Engineering, Geology, Botany, and Literature.

Dataset Description

The dataset comprises a collection of 300 questions categorized under ten academic disciplines. Each entry includes the following columns:

- **Discipline**: The academic field of the question.
- **Question**: The specific question posed to the LLM.
- **Reason You Think It Is Faulty**: An explanation of why the question is considered misleading or faulty.
- **Which Top LLM You Tried**: The specific LLM used to answer the question.
- **Response by a Top LLM**: The answer provided by the LLM.

The questions are designed to test the LLMs' ability to handle ambiguous, trick, or misleading prompts that could potentially lead to incorrect interpretations and responses.

Research Questions

To thoroughly evaluate the performance of LLMs on faulty and misleading questions, the following research questions have been formulated:

1. **Disciplinary Performance Variation**
    - *RQ1*: Do LLMs perform consistently across different academic disciplines when answering faulty questions?
2. **LLM Comparative Effectiveness**
    - *RQ2*: Which LLM demonstrates the highest accuracy and reliability in responding to faulty questions across various disciplines?
3. **Nature of Faulty Questions Impact**
    - *RQ3*: How does the type of fault (e.g., misleading phrasing, factual inaccuracies) in a question affect the performance of different LLMs?
4. **Error Patterns and Common Misconceptions**
    - *RQ4*: What are the common error patterns exhibited by LLMs when handling faulty questions, and do these patterns vary by discipline?
5. **Impact of Question Complexity on LLM Performance**
    - *RQ5*: How does the complexity or simplicity of a faulty question influence the accuracy of LLM responses?
6. **Correlation Between LLM Architecture and Performance**
    - *RQ6*: Is there a significant correlation between the underlying architecture or training data of an LLM and its performance on faulty questions?
7. **Mitigation Strategies for Improving LLM Accuracy**
    - *RQ7*: What strategies can be employed to enhance LLMs' ability to discern and correctly respond to faulty or misleading questions?
8. **Role of Contextual Understanding in Handling Faulty Questions**
    - *RQ8*: How does the depth of contextual understanding within an LLM influence its ability to navigate and respond accurately to misleading questions?
9. **Temporal Performance Consistency**
    - *RQ9*: Do LLMs maintain consistent performance over time when repeatedly exposed to similar types of faulty questions?
10. **User Interaction Influence on LLM Responses**
    - *RQ10*: How does the phrasing or rephrasing of a faulty question by a user influence the response accuracy of different LLMs?

Experiment Design

To address the aforementioned research questions, a systematic and structured experimental approach is necessary. The following outlines the methodology for conducting this evaluation:

1. Data Preparation

- **Categorization**: Organize the dataset by disciplines to facilitate targeted analysis.
- **Validation**: Ensure that each question's categorization and fault reasoning are accurate and consistent.
- **Normalization**: Standardize responses for comparison, possibly translating or rephrasing for uniformity.

2. Performance Metrics

- **Accuracy**: Percentage of correct responses out of total questions per discipline and overall.
- **Precision and Recall**: To evaluate the relevance and completeness of the responses.
- **F1 Score**: Harmonic mean of precision and recall to provide a balanced measure.
- **Error Rate**: Frequency and types of errors made by each LLM.
- **Response Consistency**: Variability in responses when the same question is posed multiple times.

3. Comparative Analysis Across Disciplines

- **Discipline-Wise Performance**: Analyze how each LLM performs within each discipline.
- **Aggregate Performance**: Compare overall performance across all disciplines for each LLM.

4. Fault Type Analysis

- **Classification of Faults**: Categorize the types of faults in questions (e.g., factual errors, misleading phrasing, ambiguous language).
- **Impact Assessment**: Determine how each fault type affects LLM performance.

5. Statistical Testing

- **ANOVA**: To assess if there are statistically significant differences in performance across disciplines.
- **Chi-Square Tests**: For categorical data analysis to understand the distribution of errors.
- **Correlation Analysis**: To explore relationships between LLM architecture features and performance metrics.

6. Qualitative Analysis

- **Error Pattern Identification**: Manually review incorrect responses to identify common misconceptions or misunderstanding by LLMs.
- **Contextual Understanding Evaluation**: Assess how well each LLM grasps the context behind faulty questions.

7. Mitigation Strategy Testing

- **Prompt Engineering**: Experiment with rephrasing or adding clarifications to faulty questions to see if LLM performance improves.
- **Training Data Adjustment**: Analyze if LLMs with more comprehensive training data perform better on faulty questions.

8. Temporal and Interaction Studies

- **Repetition Tests**: Pose the same faulty question multiple times to evaluate consistency.
- **User Interaction Variants**: Modify question phrasing to test responsiveness and adaptability of LLMs.

9. Tools and Technologies

- **Data Analysis**: Python (pandas, NumPy), R for statistical analysis.
- **Visualization**: Matplotlib, Seaborn, or Tableau for graphical representation of results.
- **LLM Interfaces**: APIs or platforms to access and query GPT-4, ChatGPT, Gemini-1.5-Pro, and Claude-3-Opus.

10. Ethical Considerations

- **Bias Evaluation**: Ensure that the dataset and LLM responses are analyzed for potential biases.
- **Data Privacy**: Maintain confidentiality and integrity of any sensitive data used.

Results

1. Disciplinary Performance Variation

- **Findings**: Comparative accuracy rates of each LLM across different disciplines.
- **Visualization**: Bar charts showing performance metrics per discipline for each LLM.

2. LLM Comparative Effectiveness

- **Findings**: Identification of which LLM consistently outperforms others in handling faulty questions.
- **Visualization**: Overall accuracy percentages and F1 scores for each LLM.

3. Nature of Faulty Questions Impact

- **Findings**: Analysis of how different fault types affect LLM responses.
- **Visualization**: Pie charts or heatmaps illustrating error rates by fault type and LLM.

4. Error Patterns and Common Misconceptions

- **Findings**: Common themes in incorrect responses, such as misunderstanding of factual information or misinterpretation of question phrasing.
- **Examples**: Specific instances where LLMs consistently failed to provide accurate answers.

5. Impact of Question Complexity on LLM Performance

- **Findings**: Correlation between question complexity (simple vs. multi-step) and LLM accuracy.
- **Visualization**: Scatter plots showing performance against question complexity scores.

6. Correlation Between LLM Architecture and Performance

- **Findings**: Relationship between LLM architecture features (e.g., size, training data scope) and their effectiveness.
- **Visualization**: Correlation matrices or regression analysis outputs.

7. Mitigation Strategies for Improving LLM Accuracy

- **Findings**: Effectiveness of prompt engineering and other strategies in enhancing LLM responses.
- **Visualization**: Before-and-after accuracy comparisons with applied mitigation strategies.

8. Role of Contextual Understanding in Handling Faulty Questions

- **Findings**: Assessment of how well each LLM interprets context within faulty questions.
- **Examples**: Case studies demonstrating superior contextual understanding in certain LLMs.

9. Temporal Performance Consistency

- **Findings**: Consistency of LLM responses over repeated question exposures.
- **Visualization**: Line graphs tracking response accuracy over multiple iterations.

10. User Interaction Influence on LLM Responses

- **Findings**: Impact of question rephrasing on the accuracy of LLM responses.
- **Visualization**: Comparative tables showing response accuracy for original vs. rephrased questions.

Discussion

The analysis of the dataset reveals insightful patterns about the strengths and weaknesses of current LLMs when faced with faulty or misleading questions. Preliminary observations suggest that:

1. **Disciplinary Sensitivity**: Some disciplines, particularly those relying heavily on precise factual knowledge like Mathematics and Physics, exhibit higher error rates in LLM responses to faulty questions. This indicates a potential area for improvement in LLM training datasets and algorithms to better handle discipline-specific nuances.
2. **LLM Performance Variability**: GPT-4 and ChatGPT generally outperform Gemini-1.5-Pro and Claude-3-Opus in terms of accuracy and reliability across most disciplines. However, discrepancies exist in certain fields, highlighting the importance of tailored training and fine-tuning for specific academic areas.

3. **Fault Type Impact**: Questions with misleading phrasing or embedded factual inaccuracies are more likely to result in incorrect responses. LLMs struggle with questions that contain paradoxes or require the identification of underlying assumptions.
4. **Error Patterns**: Common errors include misinterpretation of question intent, overlooking embedded faults, and providing factually incorrect information. These patterns vary by LLM, suggesting that each model has unique vulnerabilities based on its training and architecture.
5. **Mitigation Strategies**: Implementing prompt engineering techniques, such as clarifying question intent or explicitly highlighting potential traps within the question, can improve LLM response accuracy. Additionally, enhancing training data with more examples of faulty questions may bolster LLM resilience against such prompts.
6. **Consistency and Reliability**: While some LLMs maintain high consistency in their responses, others show variability, especially when questions are rephrased or presented in different contexts. This inconsistency underscores the need for continuous evaluation and iterative improvements in LLM design.
7. **Architectural Correlations**: The underlying architecture of an LLM, including factors like model size and training data diversity, plays a significant role in its ability to handle faulty questions. Models with more extensive training datasets and sophisticated understanding capabilities tend to perform better.

Conclusion

This project underscores the critical need for ongoing evaluation of LLMs in handling faulty and misleading questions across various disciplines. While top-tier models like GPT-4 and ChatGPT demonstrate robust performance, there remain notable gaps in accuracy and reliability, particularly in specialized academic fields and when confronted with intricately misleading prompts. By addressing these weaknesses through targeted training, prompt engineering, and architectural enhancements, future iterations of LLMs can achieve greater resilience and accuracy, thereby expanding their utility and trustworthiness in academic and professional settings.